Danbo CHEN

Regression Analysis Frequently Asked Questions

Question 1. 数据为什么要取对数

1. 缩小数据的绝对数值, 方便计算。例如, 每个数据项的值都很大, 许多这样的值进行计算可能对超过常用数据类型的取值范围, 这时取对数, 就把数值缩小了, 例如 TF-IDF 计算时, 由于在大规模语料库中, 很多词的频率是非常大的数字。

2. 取对数后, 可以将乘法计算转换称加法计算。

3. 某些情况下, 在数据的整个值域中的在不同区间的差异带来的影响不同。例如, 中文分词的 mmseg 算法, 计算语素自由度时候就取了对数, 这是因为, 如果某两个字的频率分别都是 500, 频率和为 1000, 另外两个字的频率分别为 200 和 800, 如果单纯比较频率和都是相等的, 但是取对数后,log500=2.69897, log200=2.30103, log800=2.90308 这时候前者为 2log500=5.39794, 后者为 log200+log800=5.20411, 这时前者的和更大, 取前者。因为前面两个词频率都是 500, 可见都比较常见。后面有个词频是 200, 说明不太常见, 所以选择前者。

4. 取对数之后不会改变数据的性质和相关关系, 但压缩了变量的尺度, 例如 800/200=4, 但 log800/log200=1.2616, 数据更加平稳, 也消弱了模型的共线性、异方差性等。

5. 所得到的数据易消除异方差问题。

6. 在经济学中, 常取自然对数再做回归, 这时回归方程为 lnY=a lnX+b , 两边同时对 X 求导,1/Y*(DY/DX)=a*1/X, b=(DY/DX)*(X/Y)=(DY*X)/(DX*Y)=(DY/Y)/(DX/X) 这正好是弹性的定义。

Question 2. Paraphrase the following passage. Technical or specialized terms which should not be changed are underlined :

Individual reading passages were used during baseline, intervention, and probe sessions. Reading passages utilized in this study were found on an educational website (Education.com) which offered educational tools and learning resources for parents and educators with lessons ranging from pre-kindergarten through high school.

Individual reading passages were employed during baseline, intervention, and probe sessions. These reading materials were sourced from an educational website (Education.com) that provides parents and educators with various learning tools and resources, including lessons spanning from pre-kindergarten to high school.