思考了三个方面关于缺陷预测做断点回归：

（1）对于每一个面向对象度量，超过某一个阀值，则预测该函数、类或文件有缺陷。选此阀值为断点。用传统的Logit模型研究，面向对象度量对是否存在缺陷进行预测。

但这一做法的难点在于阀值的选取，它受外界的影响因素较多，故放弃。

（2）对类规模的大小设定一个阀值，并以此为断点。当类规模大小超过这个阀值时，缺陷数量会发生变化。同（1）一样，此阀值难于选取。也放弃。

（3）把断点设定在不同的语言上。即每个面向对象度量在同一种语言内最有一个AUC值度量其预测性能，但用同一个度量在不同语言上，AUC值会发生相应的跳跃变化。即通过断点回归来预测，同一度量在不同语言上缺陷预测性能的差异。

此外，断点回归主要解决内生性问题。具体来说，影响软件缺陷（二分变量或连续变量）的因素，如耦合、继承、内聚和开发者选取语言的偏好因素中，有些是与语言选取相关，有些与语言选取无关。

用数学的语言表示，在一个简单的线性模型下：

Y=a+b\*D+e

Y：AUC值等缺陷预测性能指标。

D＝1 如果选取J语言，不使用C语言，D=0 如果选取C语言，不使用J语言。

e: 其他影响缺陷预测性能的因素。 假设 e=u+v, 其中u是指预测效果影响因素中与语言相关的因素，即和D是相关的。而v是其他影响缺陷预测性能但是和D不相关的变量。

b: 就是我们要找的“语言的选取对缺陷预测性能的影响”

论文题目：

不同语言在代码度量对软件缺陷预测性能差异的影响研究——基于语言选取断点回归分析

The Impact of Different Languages on the Performance Differences of Software Defects Prediction Based on Code Metrics——Based on Fuzzy Regression Discontinuity Analysis of Language Choice

对于某一个代码度量：



其中：

：即同一代码度量两个语言得出的某一性能指标的差值。此差值始终取绝对值。

：若PD值大于0，即i语言对j语言的差异，取1；若PD值小于0，即j语言对i语言的差异，取0；

代表了以类规模作为执行变量的RD函数形式；k的取值范围为i,j。

OO metrics

bugs

Size

Other control variables

AUC

Meta analysis

Fuzzy RD coefficient