**估计Bootstrap的confidence interval与prediction interval**

# **1）背景信息: 回归模型的预测是对目标变量的均值估计**

只有理解了以下论断，才能明白为什么基于bootstrap的N次预测值直接计算quantile interval得到的是confidence interval，而不是prediction interval:

**回归模型的预测输出，本质上是输出平均趋势，而不是单个观测值的全部可能性。**

以线性回归为例：我们假设 Y 的变化可以用一个线性关系 来描述，同时存在随机噪声，而满足i.i.d，因此均值为0：

因此，当时，Y的条件期望是：

预测值是模型根据拟合好的线性关系对目标变量的**条件均值的估计**，描述的是Y在时的平均趋势，而不是单个Y的真实值。

# **2) Confidence interval (CI) 与Prediction Interval (PI)**

首先区分两个概念：

* Confidence interval用于描述条件均值的不确定性。利用估计的回归方程，对于自变量X的一个给定值，求出因变量Y的平均值的估计区间
* Prediction interval用于描述真实观测值的不确定性（条件均值+残差）。利用估计的回归方程，对于自变量X的一个给定值，求出因变量Y的个别值的估计区间

一个十分简单的例子：掷骰子。

Confidence interval：对期望值（约3.5）的估计区间会随样本量增加而变窄。

Prediction interval：对下一次掷骰子结果的预测区间始终约为1到6，即使样本量很大也不会收窄5。

这两个不同的interval，实际上是用于两个不同的现实应用场景：

***(Confidence interval)*** 如果我们想要向投资人展示企业的商业价值，基于它过去N年的产值数据，来评估它**未来平均年产值的区间范围**（条件均值的不确定性）。在这个情境下，我们需要的是confidence interval。由于残差满足i.id，期望为0（见公式1）。此时预测model的uncertainty只有一个来源：

* 模型预测均值的不确定性

CI的公式如下：

其中，是模型预测的均值，是样本标准差，n是样本数量。

CI的范围比较窄，因为它描述的是企业“长期来看”的平均产值在哪个范围内。它提供的信息是“在未来，企业的平均年产值的范围是多少“。

**(Prediction interval)** 而在另一个场景下，我们不是关心未来长期的年产值的期望，而是关心**明年具体产值的可能性范围**。“明年的产值”，是一个具体的单点观测值，它的变化范围受到两个因素的影响：

* 模型预测均值的不确定性
* 残差（随机误差）的影响

PI的计算公式如下:

其中，是模型预测的均值，是样本标准差，是残差。被用于描述模型均值的不确定性。V被用于描述残差的不确定性。

**PI比CI的区间更宽，因为它除了考虑期望的范围，还考虑了每一个观测值的随机波动。描述的信息是“在明年，企业可能的产值范围是多少”。**

|  |  |
| --- | --- |
|  | regression - Bootstrap Prediction Interval: which residuals to use and  which method? - Cross Validated |

**图1. CI与PI的宽度差异**

如图1右所示，预测区间的范围宽于置信区间。置信区间表示真实回归线在一定置信水平下的位置（只考虑回归线的不确定性）。预测区间表示新观测值可能出现的位置，同时考虑回归线的不确定性和数据的可变性。

# **3） Bootstrap的CI与PI**

假设一个空间预测任务，测试集中共有N个样本点。目前的需求是使用bootstrap方法，对每一个样本点都得到一个预测区间prediction interval。

**Confidence interval**

假设在bootstrap中，每一次都在训练集采样M个样本点拟合回归模型，然后在测试集中进行预测，重复N次。对于测试集中的每一个样本点，都能得到N个预测值。对N个预测值从小到大排序，取5%和95%的数值，得到的范围，是confidence interval。

**Prediction interval**

因此，每次 Bootstrap 重新采样并拟合模型，对的预测值代表当前采样条件下的条件均值。如果想要得到prediction interval，必须要对每一个输出的Y，加入随机残差。

一个简单的方案是，在进行boostrap的每次采样时，都来计算**训练集上**的error，得到一个error的集合Err。然后在测试集进行预测时，得到了每一个样本点的pred，然后用pred+sample（Err）。通过这样的操作，可以给每一个样本点的输出（条件均值），加入一个可信的随机误差，从而从条件均值拓展到可能的真实值。之后的操作和计算confidence interval是一致的。在boostrap进行N次后，对于每一个测试集中的样本点，都存在N个pred+sample(Err)。对这N个value进行从小到大排序，得到5%与95%的quantile。这样得到的就是prediction interval。

# **参考资料**

* <https://www.datacamp.com/blog/confidence-intervals-vs-prediction-intervals>

从概念上讨论CI与PI的差异

* <https://olivier-roustant.fr/wp-content/uploads/2018/09/bootstrap_conf_and_pred_intervals.pdf>

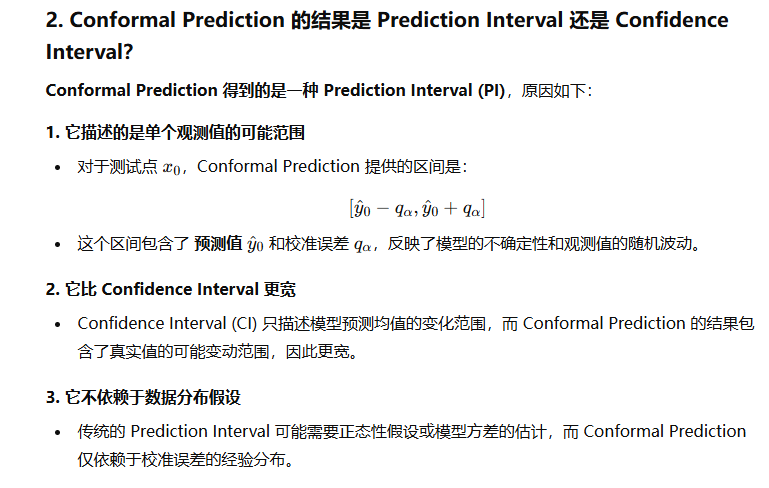
对bootstrap 的CI、PI的公式推导

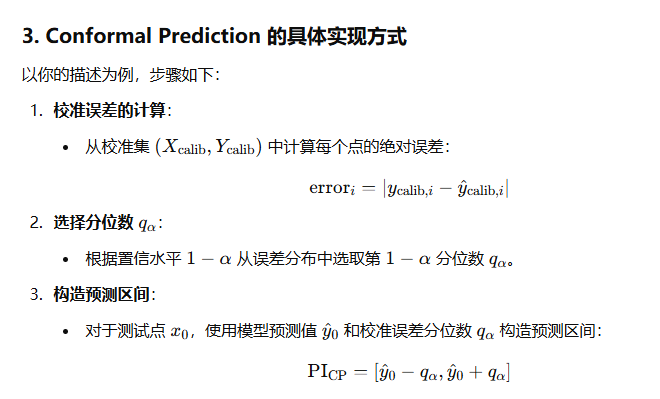
* <https://stats.stackexchange.com/questions/226565/bootstrap-prediction-interval>

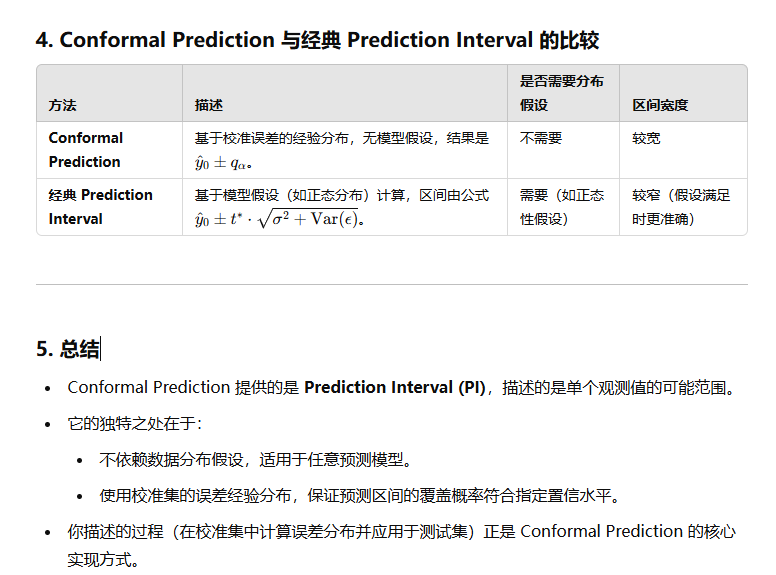
从R代码层面展现CI与PI的差异，并且提供了计算bootstrap PI的两种代码实现

* + 对代码的讲解视频：<https://www.youtube.com/watch?v=c3gD_PwsCGM&t=466s>

**补充材料：Conformal prediction得到的是Prediction interval**







**参考资料**

* <https://medium.com/bain-inside-advanced-analytics/conformal-prediction-an-easy-way-to-estimate-prediction-intervals-c0de34c47494>

讨论了Conformal prediction是对于经典的估计prediction interval（Monte Carlo Dropout, Mean Variance Estimation, Quantile Regression）的良好替代方案。