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 $\beta_0 + \beta_1 X$ 来描述，同时存在随机噪声∈，而∈满足 i.i.d，因此均值为 0：

$$E(\in) = 0$$

因此，当$X = x_0$时，Y 的条件期望是：

$$E(Y|X = x_0) = \beta_0 + \beta_1 X$$

预测值$\hat{y}$是模型根据拟合好的线性关系$\beta_0 + \beta_1 X$对目标变量的**条件均值的估计**，描述的是 Y 在$X = x_0$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：

$$CI = \hat{\mu} \pm t * \frac{\sigma}{\sqrt{n}}$$

其中，$\hat{\mu}$是模型预测的均值，σ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：

$$PI = \hat{\mu} \pm t * \sqrt{\sigma^2 + Var(\varepsilon)}$$

其中，$\hat{\mu}$是模型预测的均值，σ是样本标准差，ε是残差。σ^2被用于描述模型均值的不确定性。V$ar(\varepsilon)$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”。

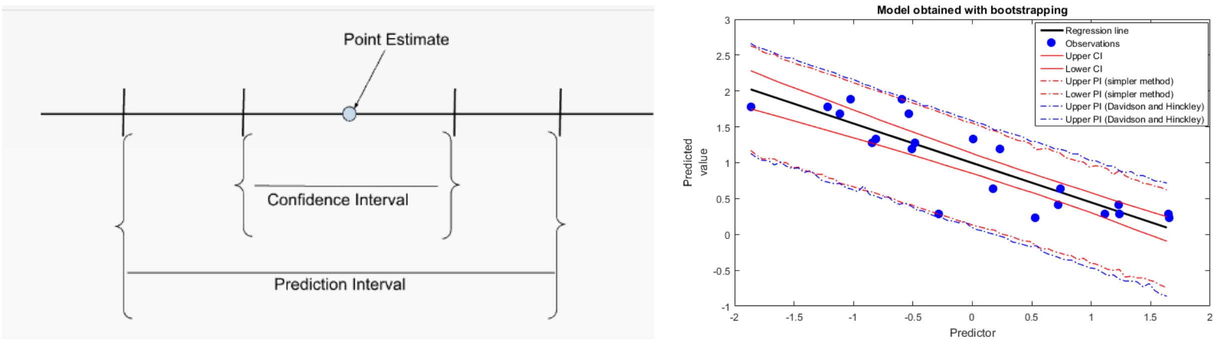

图 1. CI 与 PI 的宽度差异

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$X = x_0$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

2. Conformal Prediction 的结果是 Prediction Interval 还是 Confidence Interval?

Conformal Prediction 得到的是一种 Prediction Interval (PI)，原因如下：

1. 它描述的是单个观测值的可能范围

- 对于测试点 x_0，Conformal Prediction 提供的区间是：

$$[\hat{y}_0 - q_\alpha, \hat{y}_0 + q_\alpha]$$

- 这个区间包含了 **预测值** $\hat{y}_0$ 和校准误差 q_α，反映了模型的不确定性和观测值的随机波动。

2. 它比 Confidence Interval 更宽

- Confidence Interval (CI) 只描述模型预测均值的变化范围，而 Conformal Prediction 的结果包含了真实值的可能变动范围，因此更宽。

3. 它不依赖于数据分布假设

- 传统的 Prediction Interval 可能需要正态性假设或模型方差的估计，而 Conformal Prediction 仅依赖于校准误差的经验分布。

3. Conformal Prediction 的具体实现方式

以你的描述为例，步骤如下：

1. **校准误差的计算**：
   - 从校准集 $(X_{\text{calib}}, Y_{\text{calib}})$ 中计算每个点的绝对误差：

$$\text{error}_i = |y_{\text{calib},i} - \hat{y}_{\text{calib},i}|$$

2. **选择分位数 q_α**：
   - 根据置信水平 $1-\alpha$ 从误差分布中选取第 $1-\alpha$ 分位数 q_α。
3. **构造预测区间**：
   - 对于测试点 x_0，使用模型预测值 $\hat{y}_0$ 和校准误差分位数 q_α 构造预测区间：

$$\text{PI}_{\text{CP}} = [\hat{y}_0 - q_\alpha, \hat{y}_0 + q_\alpha]$$

4. Conformal Prediction 与经典 Prediction Interval 的比较

方法	描述	是否需要分布假设	区间宽度
Conformal Prediction	基于校准误差的经验分布，无模型假设，结果是 $\hat{y}_0 \pm q_\alpha$。	不需要	较宽
经典 Prediction Interval	基于模型假设（如正态分布）计算，区间由公式 $\hat{y}_0 \pm t^* \cdot \sqrt{\sigma^2 + \text{Var}(\epsilon)}$。	需要（如正态性假设）	较窄（假设满足时更准确）

---

5. 总结

- Conformal Prediction 提供的是 **Prediction Interval (PI)**，描述的是单个观测值的可能范围。
- 它的独特之处在于：
  - 不依赖数据分布假设，适用于任意预测模型。
  - 使用校准集的误差经验分布，保证预测区间的覆盖概率符合指定置信水平。
- 你描述的过程（在校准集中计算误差分布并应用于测试集）正是 Conformal Prediction 的核心实现方式。

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。