“Как я использовал бутстрэп?”

Автор: Бекишев Рустам Адлбекович

В рамках семинара мы использовали "bootstrap", специальный метод, благодаря которому мы можем проводить оценки статистических показателей, таких как среднее, дисперсия, доверительные интервалы и другие. Процесс осуществляется путем многократного повторения выборки из имеющегося набора данных. Этот метод является непараметрическим и основывается на использовании случайных выборок с возвращением из исходных данных.

Для построения доверительных интервалов на основе bootstrap в литературе наиболее часто используются следующие типы интервалов:

1) The Normal Interval
2) Pivotal Intervals
3) Percentile Intervals

В рамках семинара мы в основном пользовались Percentile Intervals, доказательство использования которого изложено в книге "All of statistics: a concise course in statistical inference".

Для проведения эксперимента по работе с bootstrap, я выбрал базу данных с Kaggle соревнования **House Prices - Advanced Regression Techniques.** (https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview)

Данный датасет включает в себя информацию по стоимости квартир в зависимости от большого количества характеристик, в связи с тем, что зависимая величина для построения модели является SalePrice, нам были бы интересно для нее посмотреть дополнительные характеристики.

Код для данной задачи в следующем виде (Bootstrap для таргет-переменной SalePrice):

```
import pandas as pd
import numpy as np
import seaborn as sns
data = pd.read_csv("House_Prices.csv")
target = data["SalePrice"]
!pip install arch
from arch.bootstrap import IIDBootstrap
boot_x = IIDBootstrap(target, seed=111111)

Bootstrap-интервал для среднего величины SalePrice:
boot_x.conf_int(np.mean, method='percentile', reps=10000, size=0.95)
array([[176864.26873288], [185008.99135274]])

Bootstrap-интервал для медианы величины SalePrice:
boot_x.conf_int(np.median, method='percentile', reps=10000, size=0.95)
array([[158950.], [167700.]])
```

```
Bootstrap-интервал для стандартного отклонения величины SalePrice:
boot_x.conf_int(np.std, method='percentile', reps=10000, size=0.95)
array([[73625.23379962], [85396.91639693]])


Bootstrap интервал для дисперсии величины SalePrice:
boot_x.conf_int(np.var, method='percentile', reps=10000, size=0.95)
array([[5.41246874e+09], [7.31535062e+09]])
```

Мы получили следующие результаты:

1) Bootstrap доверительный интервал для среднего SalePrice равен [176864.26873288, 185008.99135274]
2) Bootstrap доверительный интервал для медианы SalePrice равен [158950., 167700.]
3) Bootstrap-интервал для стандартного отклонения величины SalePrice равен [73625.23379962, 85396.91639693]
4) Bootstrap интервал для дисперсии величины SalePrice равен [5.41246874e+09, 7.31535062e+09]

Таким образом мы использовали бутстреп для нахождения дополнительных описательных характеристик, которые мы можем рассмотреть перед переходом к построению моделей.

Additional Reference:

Larry, Wassennan. "All of statistics: a concise course in statistical inference." (2004). (Chapter 8 - The Bootstrap)