В классе мы обсуждали кросс-валидацию и бутстраппинг. Напомню, что кросс-валидация это способ оценить истинное качество работы регрессии или классификации, а бутстрап — это способ оценки распределения чего угодно (оценок, ошибок). Бутстрап можно использовать как альтернативу кросс-валидации для проверки модели и как способ оценить распределение (обычно просто строится доверительный интервал) для оценок параметров модели. Следует помнить, что бутстрап имеет свойство занижать оценку дисперсии (т.е. бутстрап- доверительный интервал может быть уже, чем точный), а параметрический доверительный интервал в бутстрапе имеет смысл только если распределение оцениваемой величины близко к нормальному. Если распределение отлично от нормального, то следует использовать перцентильный (непараметрический) доверительный интервал.

1 Домашнее задание

Задание 1.1 (Задание (от Антона))**.** Промоделировать выборку из 50 индивидов, 1000 случайных предикторов и случайный респонз. Убедиться, что среди 1000 предикторов найдутся сильно коррелирующие с респонзом. Отобрать из них 20 самых коррелирующих, построить на них линейную регрессию и проверить ее с помощью кросс-валидации. Убедиться в справедливости утверждения о том, что полученная по CV оценка неверна, для чего исходно промоделировать выборку из 100 индивидов и рассмотреть последние 50 как тестовый набор (отбор признаков также проводить по обучающему набору).

Сделать честную кросс-валидацию — отбор признаков тоже проводить только по обучающему набору (отбор — часть обучения). Убедиться, что регрессия неинформативна.

Опционально: провести аналогичные действия для классификации. В качестве классификатора можно взять логистическую регрессию.

Задание 1.2. Повторить мой параметрический бутстраппинг ирисов для QDA-модели и QDA-классификатора. То есть надо оценить матрицы ковариаций для трех сортов независимо и промоделировать распределение из смеси многомерных нормальных. Помочь в этом вам могут функции `tapply()`, `aggregate()`, `by()` и `rmvnorm()` из пакета `mvtnorm`.

Задание 1.3. Для любой из задач регрессии (`Universities`, `concrete`, ...) с помощью бутстрапа (непараметрического) построить доверительные интервалы для коэффициентов (нужно проверить распределение оценок на нормальность и выбрать правильный способ построение д.и.).

Также с помощью кросс-валидации сравнить вашу итоговую модель с другими, убедиться, что Вы сделали правильный выбор.