

## 9. Оценки наименьших квадратов. Гауссовская линейная модель

1. Загрузите [данные](#) из набора [Forest Fires](#) о лесных пожарах в Португалии. Задача состоит в том, чтобы с помощью линейной регрессии научиться предсказывать координату area (площадь пожара) в виде линейной комбинации других данных.

*Преобразование данных.* Чтобы работать с числовыми координатами, нечисловые координаты (month, day) нужно перевести в числовые. Для простоты можно заменить координату month на индикатор летнего сезона, а координату day не использовать вообще. По желанию можете сделать преобразование другим способом. Так же желательно добавить координату, тождественно равную единице. Она будет отвечать свободному члену в линейной комбинации.

Разбейте выборку на две части в соотношении 7:3. Перед этим желательно ее перемешать (`random.shuffle`). По первой части постройте регрессионную модель. Примените модель ко второй части выборки и посчитайте по ней среднеквадратичную ошибку.

Сделайте для area преобразование  $f(x) = \ln(c+x)$  и постройте для нее регрессионную модель. Посчитайте среднеквадратичную ошибку для преобразованных значений по данному правилу и для исходных, применив в последнем случае к оценкам обратное к  $f$  преобразование. При каком  $c$  предсказания получаются лучше всего?

При выбранном  $c$  сделайте разбиение выборки в соотношении 7:3 разными способами (перемешивая каждый раз). Сильно ли зависит качество от способа разбиения? Сделайте выводы.

2. Пусть  $X_i = \beta_1 + i\beta_2 + \varepsilon_0 + \dots + \varepsilon_i, i = 0, 1, \dots, n$  — расстояния, которое проехал трамвай за  $i$  секунд по показанию датчика. Здесь  $\beta_1$  — начальное расстояние,  $\beta_2$  — скорость трамвая,  $\varepsilon_0$  — ошибка начального показания датчика. Трамвай едет с постоянной скоростью, и через каждую секунду датчик фиксирует расстояние, которое проехал трамвай. Отсчет времени идет от предыдущего замера, причем отсчет происходит с ошибкой. Для  $i = 1, \dots, n$  величина  $\varepsilon_i$  есть ошибка приращения расстояния, то есть  $\varepsilon_i = \varepsilon_i^t \beta_2$ , где  $\varepsilon_i^t$  — ошибка отсчета времени. Все ошибки  $\varepsilon_i$  независимы и распределены по закону  $N(0, \sigma^2)$ . Сведите задачу к линейной модели и найдите оценки наименьших квадратов для начального расстояния  $\beta_1$  и скорости  $\beta_2$ , а также несмещенную оценку для  $\sigma^2$ , из которой выразите оценку дисперсии отсчета времени. Данные взять из файла на диске. Сделайте выводы.

3. Сгенерируйте выборку  $X_1, \dots, X_{100}$  из стандартного нормального распределения. Постройте и визуализируйте точный доверительный интервал уровня доверия  $\gamma = 0.95$  для

- (a)  $a$  при известном  $\sigma^2$ ,
- (b)  $\sigma^2$  при известном  $a$ ,
- (c)  $a$  при известном  $\sigma^2$ ,
- (d)  $\sigma^2$  при известном  $a$ ,
- (e) доверительную область для  $(a, \sigma^2)$ .

Сделайте выводы.