Лабораторная работа №3. Регрессионный анализ. (Срок сдачи: **02.12.2021**)

Основное задание (**4 балла** + **до 1 балла** за оригинальность)

1. Смоделируйте выборку объема $N = 101$ из трехмерного нормального закона распределения. Параметры распределения должны генерироваться случайно из некоторого неконечного множества (само множество и способ генерации вы должны придумать сами (**+0,5 балла** за *оригинальность*)). Все компоненты случайного вектора должны быть коррелированны. Для генерации можно использовать любую встроенную для этих целей функцию. (**+0,5 балла**, если придумаете и реализуете «*оригинальный»* метод, использующий только генерацию одномерных нормальных случайных величин). Т.е. фактически в этом пункте 2 задания: а) генерация параметров распределения; б) генерация многомерного нормального распределения с заданными параметрами.
2. Случайным образом удалить из выборки одно трехмерное наблюдение и сохранить отдельно.
3. Обозначим компоненты сгенерированного нормального вектора как (x_1, x_2, x_3). Предположим, что имеет место модель линейной регрессии: $x_1 = a_0 + a_2x_2 + a_3x_3 + u$, где u – независимые в совокупности нормально распределенные ошибки (остатки, residuals). Оценить коэффициенты линейной регрессии a_0, a_2, a_3, математическое ожидание и дисперсию ошибки u. (Можно использовать готовые методы)
4. Для отдельно сохраненного вектора, по компонентам x_2 и x_3, используя модель линейной регрессии, спрогнозировать компоненту x_1. Найти квадрат разности точного значения этой компоненты и прогноза (квадрат ошибки).
5. Повторить шаги 1-4 $M = 100$ раз. На каждой итерации на первом шаге использовать одни и те же параметры. Насколько «хорошо» прогнозируется значение x_1. Усреднить квадрат ошибки (получим среднеквадратичную ошибку прогнозирования). Сделать выводы.

Замечание: Сама по себе среднеквадратическая ошибка прогнозирования не сильно показательна. Обычно эту величину используют для сравнения разных методов прогнозирования или для изучения поведения прогноза в зависимости от изменения каких-либо параметров.

Дополнительные задания

Разное

6. (**+1 балл**) Найти истинные коэффициенты регрессии модели из пункта 3, сравнить полученные результаты.
7. (**+0,5 балла**) Предположим, что имеет место модель линейной регрессии: $x_1 = a_2x_2 + a_3x_3 + u$, где u – независимые в совокупности нормально распределенные ошибки (остатки, residuals). Оценить коэффициенты линейной регрессии a_2, a_3, математическое ожидание и дисперсию ошибки u. Сравнить полученные результаты с результатами пункта 3. Сравнить среднеквадратическую ошибку прогнозирования в этом случае с ошибкой из пункта 5.

Графики

8. (**+1 балл**) Сгенерировать выборку из пункта 1. На графике изобразить диаграмму рассеяния компонент x_1 и x_2. Предположим, что имеет место модель линейной регрессии: $x_1 = a_0 + a_2x_2 + u$. Оценить коэффициенты данной модели. Построить на графике линию регрессии $x_1 = a_0 + a_2x_2$.
9. (**+0,5 балла**) Для остатков модели из пункта 8 (или пункта 3) построить q-q plot. Разобраться, что на нем изображается и дать содержательную интерпретацию того, что получилось в вашем случае.

Хи-квадрат критерий Пирсона

10. (**+2 балл**) *Реализовать* хи-квадрат критерий Пирсона для проверки гипотезы о том, что остатки модели из пункта 3 имеют нормальный закон распределения. Из встроенных функций разрешается использовать только функции распределения (cdf), плотности (pdf) и функции для вычисления квантилей различных законов распределений (quantiles).
11. (**+0,5 балла**) Сравнить полученные результаты с результатами работы встроенного теста.
12. (**+0,5 балла**) Определить, для каких значений k числа интервалов нулевая гипотеза принимается, для каких отвергается (k изменять от 3 до 50).

Переобучение (over fitting)

13. (**+1 балл**). Сгенерировать M выборок из двумерного нормального закона распределения объема N с нулевым математическим ожиданием и недиагональной матрицей ковариации. В каждой выборке случайным образом выбрать 10% наблюдений, которые исключаются из выборки и хранятся отдельно (проверочные выборки). По оставшимся наблюдениям оценить коэффициенты регрессии $x_1 = a_0 + a_1 x_2 + a_2 x_2^2 + ... + a_p x_2^p + u$ для фиксированного параметра p. Спрогнозировать значения x_1 для проверочных выборок и найти среднеквадратическую ошибку прогнозирования (усреднение идет по объему проверочной выборки и по M). Изучить, как изменяется среднеквадратическая ошибка прогнозирования в зависимости от p.

14. (**+0,5 балла**) Для первой выборки построить несколько диаграмм рассеяния. На них изобразить линии регрессии при различных значениях p.