Практика 5. Регрессия. МНК и свойства МНК-оценок

Часть 1. Критерий МНК как функция коэффициентов регрессии в трехмерном пространстве

1. Сгенерировать выборку объема *n* по модели **из своего варианта на сайте. Вариант выбирается по списку на сайте**. Для модели заданы истинные коэффициенты регрессии и дисперсия шума. Объем выборки *n* задать от 10 до 20.
2. Рассчитать МНК-оценки коэффициентов регрессии на основе аналитического решения задачи минимизации функционала качества. Расчет сделать функцией regress и по формуле:
$$\hat{\beta} = (X^T X)^{-1} X^T Y$$
3. Построить трехмерный график (surf) и график изолиний (contour) функционала качества МНК $Q(\beta)$. Отобразить точкой аналитическое решение (plot3). Убедиться, что точка лежит точно в минимуме.

Часть 2. Статистические свойства МНК-оценок

1. Построить распределение оценок коэффициентов регрессии как случайных величин:
    a. теоретическую плотность распределения
    b. методом статистического моделирования Монте-Карло
    c. по результатам статистического моделирования определить 95% доверительный интервал коэффициентов, вывести на график
2. Построить распределения **прогноза** $\hat{y}(x_{\text{н}})$ как случайной величины при x=200 и x = 240
    a. теоретическую плотность распределения
    b. методом статистического моделирования Монте-Карло
    c. на графиках отметить истинное значение $y_{\text{ист}}(x_{\text{н}})$
    d. по результатам статистического моделирования определить 95% доверительный интервал коэффициентов, вывести на график
3. Построить распределение ошибки прогноза **истинного** значения $e_{\text{и}} = \hat{y}(x_{\text{н}}) - y(x_{\text{н}})$ при $x_{\text{н}} = 200, x_{\text{н}} = 300$.
    a. теоретическую плотность распределения
    b. методом статистического моделирования Монте-Карло
    c. по результатам статистического моделирования определить 95% доверительный интервал коэффициентов, вывести на график
4. Построить распределения ошибки предсказания нового значения $e_{\text{н}} = \hat{y}(x) - y_{\text{н}}$ при $x_{\text{н}} = 200, x_{\text{н}} = 300$.
    a. теоретическую плотность распределения
    b. методом статистического моделирования Монте-Карло
    c. сопоставить распределения $e_{\text{и}}, e_{\text{н}}$, визуально проверить соответствие дисперсий
    d. по результатам статистического моделирования определить 95% доверительный интервал коэффициентов, вывести на график
5. Методом статистического моделирования построить полигон относительных частот ошибки предсказания истинного значения $e_{\text{и}} = \hat{y}(x) - y(x)$ при x=200 и x = 300.
6. Методом статистического моделирования построить полигон относительных частот распределения ошибки предсказания нового значения $e_{\text{н}} = \hat{y}(x) - y_{\text{н}}$ при x=200 и x = 300.
7. Построить теоретические плотности распределения $\hat{y}(x_{\text{н}}), e_{\text{и}}, e_{\text{н}}$ при всех x. Отметить 95%-доверительный диапазон. График совместить с графиками из предыдущего пункта.

Вопросы к защите

1. Объяснение невязок между моделью и данными как реализаций случайного шума.
    – помимо примера с QH-характеристикой привести свой пример

- пояснить содержательный смысл дисперсии шума в вашем примере

2. Вероятностная модель выборки до измерений (векторный вид, скалярный вид).
3. Как сформировать вектор факторов
    a. для задания линейной зависимости
    b. для задания полиномиальной зависимости
4. Как формируется матрица плана?
5. Формальная постановка задачи поиска оценок коэффициентов регрессии как задачи минимизации функционала (в векторном виде).
6. Что такое система нормальных уравнений? Как она выводится?
7. Оценки коэффициентов регрессии $\hat{\beta}$ как случайные величины.
    a. как проявляется их случайность?
    b. модель $\hat{\beta}$ до измерений
    c. доказательство линейности связи $\hat{\beta}$ и вектора шума
    d. доказательство несмещенности оценок $\hat{\beta}$
    e. вывод ковариационной матрицы оценок $cov(\hat{\beta})$
    f. какое теоретическое распределение имеет коэффициент $\hat{\beta}_i$?
    g. от чего зависит точность (дисперсия) оценивания $\hat{\beta}$? Внести изменения в коде, чтобы проиллюстрировать влияние этих факторов
8. Прогноз регрессионной модели как случайная величина
    a. ошибка прогноза истинного значения и ошибка прогноза нового значения
    b. почему важна коррелированность компонентов $\hat{\beta}$?
        i. вывод формулы дисперсии суммы коррелированных величин $D(X+Y)$
        ii. обобщение на случай $D(aX+bY)$
        iii. что играет роль a, b в прогнозе $\hat{y}(x)$? что играет роль X, Y?
    c. чем обусловлена погрешность прогноза истинного значения и погрешность нового значения?
    d. доказательство несмещенности прогноза, вывод дисперсии прогноза,
    e. дисперсии ошибок прогноза истинного значения и нового значения
    f. доверительный интервал прогноза при **известной дисперсии**
        i. прогноз истинного значения
        ii. прогноз нового значения