Практика 10. Статистические свойства оценок коэффициентов и прогноза регрессионных моделей

Часть 1. Вероятностная модель процесса и код имитации эксперимента

1. Задать вероятностную модель процесса. Для этого
   a. Взять оценки коэффициентов для модели 2-го порядка из практики, посвященной построению QH-характеристик. **В данной работе считать их истинными** ($\beta_{\text{ист}}$).
   b. Задаться дисперсией шума D_ε.
   c. Задаться объемом выборки n от 10 до 20.
   d. Задаться диапазоном значений расхода в планируемом эксперименте [Qmin, Qmax]
2. Реализовать формирователь плана эксперимента по расходу
   function Qplan = build_plan(n, Qmin, Qmax).
   Qplan – планируемые в эксперименте значения расхода, лежащие в диапазоне [Qmin, Qmax]
   В данной работе будет многократно повторяться эксперимент по одному плану, поэтому при формировании матрицы плана не допускается использование случайных генераторов. При повторном вызове функция build_plan должна возвращать такой же план, как изначально.
3. Реализовать формирователь матрицы плана – вспомогательная функция, формирующая матрицу плана для произвольного порядка вектора факторов
   function X = build_plan_matrix(Qplan, k)
   в дальнейшем в коде матрица плана должна формироваться этой функцией
4. Реализовать генератор экспериментальных данных согласно плана эксперимента
   function [Y, X] = generate_process_data(Qplan, $\beta_{\text{ист}}$, D_ε)
   Y – значение моделируемого параметра процесса
   X – матрица плана проведенного эксперимента
5. Сгенерировать выборку объемом n с помощью функции generate_process_data. По полученным данным построить оценку параметров модели $\hat{\beta}$ и вывести их в консоль. В одних осях построить:
   a. график истинной зависимости $y_{\text{ист}} = x^T \beta_{\text{ист}}$в диапазоне [0.8 Qmin, 1.2 Qmax]
   b. точки экспериментальной выборки
   c. график восстановленной зависимости $\hat{y} = x^T \hat{\beta}$ в диапазоне [0.8 Qmin, 1.2 Qmax]
6. Несколько раз запустить код, убедиться, что значения $\hat{\beta}$ изменяются и вместе с ним меняется график $\hat{y} = x^T \hat{\beta}$.

Часть 2. Статистические свойства МНК-оценок

1. Построить распределение оценок **каждого** коэффициента регрессии $\hat{\beta}_i$ 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 для $x_{\text{н}}$, сформированных при Q = 0.8 Qmin, Q = 1.2 Qmax.
   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 $e_{\text{и}} = \hat{y}(x_{\text{н}}) - y_{\text{ист}}(x_{\text{н}})$ для $x_{\text{н}}$, сформированных при Q = 0.8 Qmin, Q = 1.2 Qmax.
   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 $e_н = \hat{y}(x_н) - y_н$ для $x_н$, сформированных при Q = 0.8 Qmin, Q = 1.2 Qmax.
    a. теоретическую плотность распределения
    b. методом статистического моделирования Монте-Карло
    c. сопоставить распределения $e_и, e_н$, визуально проверить соответствие дисперсий
    d. по результатам статистического моделирования определить 95% доверительный интервал коэффициентов, вывести на график
5. Количество реализация Монте-Карло выбрать так, чтобы графики hist_density были достаточно гладкими.
6. Убедиться, что нет дублирования кода функций первой части задания.

Вопросы к защите
1. Формула МНК-оценок без вывода
2. Объяснение невязок между моделью и данными как реализаций случайного шума.
3. Вероятностная модель выборки до измерений (через матрицу плана, через вектор факторов векторный вид, скалярный вид).
4. 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 $cov(\hat{\beta})$; почему недостаточно посчитать только дисперсии.
    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
5. Прогноз регрессионной модели как случайная величина
    a. Ошибка прогноза истинного значения и ошибка прогноза нового значения. Чем обусловлена погрешность прогноза истинного значения и погрешность прогноза нового значения?
    b. Доказательство несмещенности прогноза.
    c. Вывод дисперсии прогноза. Влияние коррелированности компонентов $\hat{\beta}$ при расчете дисперсии прогноза
        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?
        iv. почему важна коррелированность компонентов $\hat{\beta}$? что будет если ее не учесть? в какой момент это учитывается?
    d. дисперсии ошибок прогноза истинного значения и нового значения
    e. доверительный интервал прогноза при **известной дисперсии**
        i. вывод правила трех сигм
        ii. почему важна несмещенность прогноза?
        iii. прогноз истинного значения
        iv. прогноз нового значения