

## Практика 13. Регрессионные модели характеристик насосных агрегатов

(начало совпадает с практикой 4, новое начиная с части 3)

### Часть 1. Подготовка данных

1. Скачать с сайта курса данные по характеристикам насосных агрегатов (НА).

Данные представляют собой реальные обезличенные данные по напорным характеристикам (QH) и характеристикам мощности ( $\eta$ ) насосных агрегатов нефтеперекачивающих станций. В одном файле содержатся данные по одному НА. В имени файла указывается номер нефтеперекачивающей станции (НПС) и номер агрегата. Каждый файл содержит 3 колонки Q, H,  $\eta$ .

**Если данные по всем вариантам не выложены, нужно взять пример данных (он точно выложен) и начать отлаживать код на нем.**

2. Для зачитки данных по агрегатам воспользуйтесь функцией (есть в архиве с данными):

```
function [Q, H, N] = read_pump_data(station_number, pump_number)
```

функция на вход принимает номер НПС и номер НА (возьмите их своего варианта), на выходе выдает Q, H,  $\eta$ .

**Если данные не выложены, временно берите `station_number = 1` и `pump_number = 1`. По готовности данных, переделать под свой вариант.**

### Часть 2. QH-характеристика и характеристика КПД

1. Построить регрессионную модель с полиномом 0-го, 1-го, 2-го, 3-го порядков.
2. Построить модели на основе аппроксимации из литературы по гидравлике:
  - a. Для QH-характеристики

$$H(Q) = a - bQ^2$$

- b. Для характеристики мощности

$$\eta(Q) = k_1Q - k_2Q^2$$

**Для оценок коэффициентов моделей использовать как формулу  $\hat{\beta} = (X^T X)^{-1} X^T Y$ , так и функцию `regress`**

3. Построить график  $H(Q)$ , на который вывести
  - a. исходные данные (вывести отдельными точками, например, кружками)
  - b. предсказания всех моделей из части 2 в диапазоне Q от нуля до  $1.5 * \max(Q)$ , (график вывести линиями)
  - c. легенду для графиков всех моделей (чтобы не путать модели друг с другом)

Отмасштабировать график по H в диапазоне от нуля до  $\max(H)$ , используя функцию `ylim`.

4. Построить график  $\eta(Q)$ , на который вывести
  - a. исходные данные (вывести отдельными точками, например, кружками)
  - b. предсказания всех моделей из части 2 в диапазоне Q от нуля до  $1.5 * \max(Q)$ , (график вывести линиями)
  - c. задать диапазон по кпд. от нуля
  - d. вывести легенду для графиков всех моделей (чтобы не путать модели друг с другом)

Отмасштабировать график по  $\eta$  в диапазоне от нуля до  $\max(\eta)$ , используя функцию `ylim`.

5. Для каждой модели рассчитать вектор регрессионных остатков  $e = Y - X\hat{\beta}$ 
  - a. построить график регрессионных остатков как функции от расхода Q (вывести отдельными точками, например, кружками)
  - b. построить график регрессионных остатков по номерам экспериментов
  - c. вывести гистограмму регрессионных остатков

Провести визуальный анализ графика регрессионных остатков и их гистограмм. Найти подозрительные на ваш взгляд особенности. По возможности обосновать свои подозрения.

### Часть 3. Метрики качества модели

1. Реализовать функцию расчета коэффициента детерминации и сопутствующие функции:

```
function S2 = calc_residual_dispersion(y, y_prediction, k)
```

```
function TSS = calc_TSS(y)
```

```
function ESS = calc_ESS(y, y_prediction)
```

```
function RSS = calc_RSS(y, y_prediction)
```

```
function R2 = calc_R2(y, y_prediction)
```

y – вектор фактических значений

y\_prediction – вектор откликов тестируемой модели

k – порядок тестируемой модели

2. Рассчитать и свести в таблицу следующие показатели по всем моделям
  - a. Несмещенную оценку дисперсии шума
  - b. Коэффициент детерминации  $R^2$  с помощью calc\_R2 и с помощью функции regress. Убедиться, что значения совпадают (при несовпадении найти ошибку в Матлабе или у себя в коде).
3. Провести анализ таблицы, выбрать лучшую модель.

### Часть 4. Анализ регрессионных остатков

1. Реализовать функцию расчета статистики и функцию проверки гипотезы критерия Вальда-Вольфовица

```
function Ns = get_series_count(e)
```

```
function [Nplus, Nminus] = get_signed_values_count(e)
```

```
function [random_residuals, z, Ns] = test_residuals_randomness(e)
```

e – вектор остатков, отсортированных по величине фактора

Ns – количество серий

Nplus, Nminus – соответственно, количество положительных и отрицательных элементов в выборке

random\_residuals – равно единице, если остатки случайны, нулю если нет

z – стандартизированное значение статистики критерия

**Примечание.** При реализации корректно считайте пороги с учетом того, должен ли быть критерий односторонний или двусторонний.

2. Убедиться, что z-статистика совпадает со значением, возвращаемой функцией runstest.
3. Построить графики регрессионных остатков с сортировкой по величине возрастания фактора (расхода), по оси абсцисс – номер эксперимента в отсортированной последовательности. (Можно исправить вывод графика в части 2).
4. Рассчитать и свести в таблицу результат анализа регрессионных остатков по всем моделям (QH и КПД).
5. Выявить наиболее простую модель (QH и КПД), которая прошла тест на случайность остатков.

### Часть 5. Проверка значимости модели

1. Реализовать расчет статистики Фишера, используя реализованный ранее расчет ESS и RSS в виде отдельной функции. Реализовать функцию проверки гипотезы о значимости модели.

```
function [F, n1, n2] = model_Fstat(y, y_prediction, k)
```

```
function significant_model = test_model_significance(F, n1, n2, alpha)
```

y, y\_prediction, k – аналогичны с расчетом  $R^2$

F – значение формируемой F-статистики при проверке значимости модели, n1, n2 – степени свободы распределения данной статистики

significant\_model – равна нулю, если модель незначима, единице если значима

alpha – уровень значимости

2. Убедиться, что расчетное значение F-статистики совпадает с расчетом из функции regress.
3. Рассчитать и свести в таблицу (см. ниже) результат проверки значимости по всем моделям (QH и КПД).

### Часть 6. Проверка значимости оценок коэффициентов модели

1. Реализовать функцию расчета студентизированных оценок коэффициентов регрессии. Реализовать функцию проверки значимости коэффициентов регрессии.

```
function [beta_st, beta_bounds] = ...
    studentize_beta_est(beta_est, K_est, n, alpha)
function significant_coeff = ...
    test_coefficient_significance(beta_st, n, alpha)
```

beta\_est – оценки коэффициентов

beta\_st – студентизированные оценки коэффициентов

beta\_bounds – доверительный интервал оценок коэффициентов регрессии

$$\hat{\beta}_i - t_{\alpha} < \beta_i < \hat{\beta}_i + t_{\alpha}$$

K\_est – оценка ковариационной матрицы оценок коэффициентов

n – объем выборки, по которой строились оценки коэффициентов

significance\_coeff – вектор из k булевых значений, которые равны единице, если коэффициент значим, иначе равны нулю

alpha – уровень значимости при расчете доверительного интервала и порогов критерия

2. Убедиться, что расчет доверительных интервалов совпадает с расчетом функции regress (выходной аргумент bint [b, bint] = regress(y, X, alpha)).

3. Реализовать альтернативный способ проверки значимости коэффициентов: если ноль лежит внутри доверительного интервал оценки коэффициента, то коэффициент незначим.

```
function significant_coeff = ...
    test_coefficient_significance_alt(beta_bounds)
```

4. Рассчитать и свести в таблицу (см. ниже) результат проверки значимости всех коэффициентов по всем моделям (QH и КПД).

### Результаты практики в табличном виде

Запись результатов расчета выполнить в формате Microsoft Excel с помощью функции xlswrite.

Имя файла указывать в виде models\_ <фамилия>.xls, чтобы не было путаницы.

Коэффициенты модели	$R^2$	Анализ остатков	Оценка дисперсии шума	Значимость модели	Значимость коэффициентов (от $\hat{\beta}_0$ до $\hat{\beta}_{k-1}$ )

### Вопросы к защите

1. Повторение сведений из прошлых практик
  - a. Вероятностная модель выборки до измерений для регрессионной модели
  - b. В чем разница между  $\hat{\beta}$ ,  $\beta$ ,  $\beta_{ист}$ ?
2. Коэффициент детерминации  $R^2$ 
  - a. Тривиальная модель
    - i. Понятие тривиальной (опорной) модели. Зачем она нужна?
    - ii. Записать вероятностную модель выборки до измерений в предположении тривиальной модели
    - iii. Эквивалентность МНК-оценок тривиальной модели и прогноза по средневывборочному.
  - b. Формулы и содержательный смысл  $RSS, TSS, ESS$
  - c. Доказательство равенства  $TSS = RSS + ESS$
  - d. Формула коэффициента детерминации  $R^2$ .

- i. Чему соответствуют значения  $R^2$ , равные 0%, 50%, 100%.
  - ii. При каком  $R^2$  модель пройдет через все точки выборки?
- e. Как на основе  $R^2$  выбрать лучшую модель из нескольких кандидатов?
  - i. Какие имеются негативные следствия из того, что  $R^2$  – случайная величина? Пусть для нескольких моделей  $R^2$  отличается мало. Корректно ли выбирать модель с формально меньшим  $R^2$ .
  - ii. Какое значение примет  $R^2$  при переподгонке?
- 3. Статистический критерий серий (Вальда-Вольфовица)
  - a. Нарисовать эскизы типичных примеров рядов: случайный ряд, ряд с трендом, ряд с периодичностью. Исходя из определения серии, указать, для какого типа ряда их количество  $N_S$  больше или меньше.
  - b. Формулировки основной и альтернативной гипотезы.
  - c. Эскизы распределений статистики  $N_S$  в условиях основной и альтернативной гипотез.
  - d. Эскизы распределений статистики  $z$  в условиях основной и альтернативной гипотез.
  - e. Формула стандартизированной статистики  $z$ . Как стандартизировать нормальную СВ? От каких численных параметров зависит МО и дисперсия  $N_S$ ? Распределения статистики  $N_S$  и стандартизированной статистики  $z$  в условиях основной гипотезы (точная форма распределения).
  - f. Как построить область принятия гипотезы для статистик  $N_S, z$ ?
  - g. Является ли распределение  $N_S, z$  непрерывным или дискретным? Является ли распределение статистик строго нормальным? Какие негативные следствия при проверке гипотез из того, что распределение  $N_S, z$  аппроксимируется нормальным?
- 4. Применение критерия серий для анализа регрессионных остатков
  - a. Типичный вид графика регрессионных остатков при заниженном порядке аппроксимации. Применение критерия серий к анализу коррелированности регрессионных остатков. Какой тип ряда с точки зрения критерия серий формируется при занижении порядка аппроксимации?
  - b. Как нужно подготовить регрессионные остатки для применения критерия серий?
  - c. Чему соответствует тренд и колебательность регрессионных остатков?
  - d. Доказательство смещенности МНК-оценок при отбросе значимых факторов.
  - e. Какие из статистических свойств нарушаются при занижении порядка аппроксимации?
  - f. Какую модель следует выбрать, исходя из анализа регрессионных остатков.
- 5. Критерий Фишера
  - a. Распределение Фишера, связь с распределением хи-квадрат.
  - b. Проверка равенства дисперсий двух выборок.
    - i. Формулировка основной, альтернативной гипотезы.
    - ii. Как проверить равенство случайных погрешностей двух датчиков?
    - iii. Эскиз статистики Фишера, если дисперсии выборок равны, если первая дисперсия больше, если вторая дисперсия больше.
    - iv. Расчет области принятия основной гипотезы.
  - c. Проверка значимости модели
    - i. Что такое значимость модели? Почему возникает проблема различения значимой и незначимой моделей?
    - ii. Статистика Фишера для проверки значимости модели.
    - iii. Формулировка основной и альтернативной гипотезы.
    - iv. Эскиз графиков распределений статистики в условиях основной и альтернативной гипотезы?

- v. Какие значения статистики соответствуют значимой модели, а какие – незначимой?
    - vi. Как рассчитать пороги для области принятия гипотезы?
  - d. Почему нельзя применять критерий Фишера (и другие критерии) к модели, которая не прошла тест случайности регрессионных остатков?
  - e. Корректно ли применять критерий Фишера для проверки значимости модели, у которой регрессионные остатки не коррелированы?
6. Критерий Стьюдента для проверки значимости коэффициентов модели.
- a. Понятие значимости коэффициента.
    - i. Значимость коэффициента – имеется ввиду коэффициент из  $\hat{\beta}$  или  $\beta_{\text{ист}}$ ?
    - ii. Почему возникает проблема различения значимого и незначимого коэффициента?
  - b. Статистические свойства МНК-оценок, в том числе при использовании лишних факторов.
    - i. Распределение оценок коэффициентов регрессии – МО, ковариационная матрица. Доказательство нормальности.
    - ii. Статистические свойства МНК-оценок при использовании лишних факторов в модели
  - c. Статистика критерия значимости коэффициента регрессии
    - i. Распределение Стьюдента – формула, вид распределения. Отличие от стандартного нормального распределения.
    - ii. Формулировка основной и альтернативной гипотезы.
    - iii. Точный вид распределения в условиях основной гипотезы.
    - iv. Эскизы распределения статистики в условиях основной и альтернативной гипотезы. Содержательная интерпретация возможных значений статистики критерия.
    - v. Расчет области принятия гипотезы.
  - d. Может ли встретиться ситуация, когда:
    - i. все коэффициенты модели незначимы по t-статистике, но модель в целом значима в F-статистике?
    - ii. модель в целом незначима по F-статистике, но хотя бы часть коэффициентов значима по t-статистике?
7. Провести анализ результатов выполненных статистических тестов из таблицы.
- a. Указать корректные и некорректные модели с точки зрения регрессионного анализа.