Практическая работа № 12. Регрессионный анализ: линейная по параметрам регрессионная модель общего вида

12.1. О содержании и задачах практической работы

В этой практической работе рассматривается та же проблема, что и в практикуме № 11. Только теперь нас интересует ситуация, при которой расположение точек на диаграмме рассеивания явно указывает на нелинейную зависимость или же гипотеза о линейности отвергнута при проверке значимости коэффициента корреляции. В этом случае имеет смысл описывать связь между переменными другими, нелинейными, формулами. Здесь имеется в виду нелинейность по переменной x; по параметрам зависимость останется линейной (в этом случае говорят о линейной регрессионной модели общего вида, или о криволинейной регрессии).

Построение и исследование регрессионной модели целесообразно разбить на этапы.

На первом, предварительном, этапе обычно анализируют расположение точек на диаграмме рассеивания; это помогает избежать очевидных ошибок в выборе формы зависимости от x. Строго говоря, в рамках заданий настоящей практической работы не запланировано обсуждение проблемы выбора наилучшей формы функциональной зависимости от x, но качественный анализ визуализированных данных поможет нам при решении задачи придерживаться здравого смысла.

На втором этапе методом наименьших квадратов находят оценки параметров регрессионной кривой. Для поиска оценок параметров рекомендуется использовать матричный подход.

На третьем этапе получают интервальные оценки неизвестных параметров модели, проверяют статистические гипотезы относительно их истинных значений.

На четвертом этапе изучают вопрос об адекватности модели (ее согласованности с результатами наблюдений).

12.2. Математические понятия и утверждения: краткая информация и ссылки на источники

1. Регрессионный анализ (общий взгляд) [1, с. 207 – 219].

2. Линейная регрессионная модель общего вида (криволинейная регрессия). МНК-оценки параметров модели [2, с. 311 – 315].

3. Оценка дисперсии ошибок наблюдения. Оценка ковариационной матрицы МНК-оценок параметров. Доверительные интервалы для параметров регрессии, для дисперсии ошибок наблюдений. Проверка значимости линейной регрессионной модели общего вида [2, с. 316 – 319].

12.3. Библиотечные инструменты языка программирования Python

Загрузка основных модулей

```
import numpy as np
import scipy.stats as sts
import matplotlib.pyplot as plt
import pandas as pd
import seaborn
from sklearn.linear_model import LinearRegression
%matplotlib inline
```

Средства полиномиальной регрессии библиотеки numpy

Функция `np.polyfit(x, y, deg)`

Вычисляет оценки b_0, b_1, ..., b_{deg} коэффициентов многочлена степени `deg` методом наименьших квадратов.

Параметры: `x,y` – массивы; `deg` – степень многочлена.

12.4. Пример для совместного обсуждения

Пример 1. Для понимания методики получения оценок параметров регрессии и последующей проверки качества аппроксимации решим сформулированную ниже задачу матричным методом, почти вручную, ограничив использование языка программирования Python арифметическими вычислениями.

В таблице приведены результаты наблюдений:

x	10	10	10	10	10	20	20	20	20	35	35	35	35	35
y	5	6	5	6	7	12	13	14	13	17	19	16	15	15

x	40	40	40	40	40	60	60	60	60	60
y	18	20	21	18	20	17	19	16	14	16

Считая, что зависимость между переменными x и y имеет вид $y = \beta_0 + \beta_1 x + \beta_2 x^2$, выполним пошагово следующие действия:

1) построим диаграмму рассеивания (убедимся, что ее вид не противоречит выбранной нами модели функциональной зависимости);

2) найдем оценки параметров, выпишем уравнение регрессии, нанесем регрессионную прямую на диаграмму рассеивания;

3) вычислим оценку s^2 для дисперсии ошибок наблюдений σ^2;

4) найдем коэффициент детерминации R^2;

5) построим доверительные интервалы для параметров регрессии;

5) построим доверительные интервалы для параметров модели.

6) построим график остатков

7) проверим адекватность модели, используя подходящий критерий.

Полученные результаты интерпретируем

Пример 2.

В таблице приведены результаты наблюдений:

x	1	2	3	4	5	6	7	8	9	10
y	16,5	13,75	13,31	12,5	13,52	12,75	12,3	12,83	12,28	12,34

1) построим диаграмму рассеивания и выберем модель регрессии Y на x;

2) найдем оценки параметров выбранной модели; выпишем уравнение регрессии; нанесем регрессионную кривую на диаграмму рассеивания;

3) вычислим остаточную дисперсию;

4) проверим значимость модели на уровне $\alpha = 0,05$;

5) построим доверительные интервалы для параметров модели и дисперсии ошибок наблюдений;

6) построим график остатков (качественно оценим адекватность модели).

12.5. Задания для самостоятельного выполнения

В файле «Данные 12_1» приведены 30 двумерных выборок непрерывных случайных векторов. Выполните задания 1 и 2 для двумерной выборки, отобранной в соответствии с вашим вариантом.

Задание 1.

1) Постройте диаграмму рассеивания и проанализируйте ее с точки зрения наличия и характера связи между компонентами выборочного вектора. Оправдано ли для описания зависимости использовать модель $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + E$? Если вы считаете, что не оправдано, то для выполнения следующих заданий используйте иную линейную регрессионную модель общего вида (но обязательно отличную от $y = \beta_0 + \beta_1 x$), и далее корректируйте действия п. 2) – 6) в соответствии с выбранной вами моделью.

2) Считая, что $M[Y \mid x] = \beta_0 + \beta_1 x + \beta_2 x^2$, найдите оценки параметров модели $y = \beta_0 + \beta_1 x + \beta_2 x^2$ (непосредственно по формулам, без использования специализированных функций языка программирования Python).

3) Нанесите график построенного в п. 2 уравнения регрессии на диаграмму рассеивания.

4) В предположении, что ошибки наблюдений не коррелированы и имеют нормальное распределение $N(0, \sigma)$, оцените качество аппроксимации результатов наблюдения уравнением регрессии $y = \beta_0 + \beta_1 x + \beta_2 x^2$:

а) проверьте значимость модели на уровне $\alpha = 0,05$;

б) найдите точечные оценки дисперсии ошибок наблюдений и ковариационной матрицы;

в) определите доверительные интервалы для параметров модели и дисперсии ошибок наблюдений при уровне значимости $\alpha = 0,05$.

5) Для изучения вопроса об адекватности построенной модели проанализируйте остатки (выборку значений случайных ошибок наблюдений – разностей между наблюдаемыми значениями y_i и вычисленными

по регрессионному уравнению $\widetilde{y_i}$, $i = 1,2,...,n$). Постройте график зависимости остатков от x_j, постройте гистограмму выборки значений случайных ошибок наблюдений, проверьте гипотезу о распределении ошибок наблюдений по нормальному закону.

Замечание. При проведении регрессионного анализа считают, что случайные ошибки наблюдений имеют нулевое математическое ожидание, одинаковую дисперсию, попарно некоррелированы и распределены по нормальному закону (и, следовательно, являются независимыми случайными величинами). Подтверждение перечисленных свойств остатков говорит в пользу правильности построенной модели.

6) Сгруппируйте данные по x (данные группировки по x выведите на печать) и проверьте адекватность линейной регрессии Y на x на уровне значимости 0,05, используя стандартную методику.

Задание 2. 1) Используйте для описания статистической зависимости компонент того же выборочного вектора модель $y = \beta_0 + \beta_1 x$ (выполните действия, аналогичные перечисленным в задании 1).

2) Сопоставьте характеристики построенной регрессионной модели с характеристиками регрессионной модели задания 1. Можно ли утверждать, что какая-то из моделей предпочтительней в использовании?