Практическая работа № 11.
Регрессионный анализ:
парная линейная регрессия

11.1. О содержании и задачах практической работы

В настоящей практической работе рассматривается проблема определения формы связи между двумя наблюдаемыми в эксперименте переменными. Считается, что одна из переменных (например, x) находится под контролем экспериментатора и может быть измерена с пренебрежимо малой ошибкой, в то время как измеряемые значения другой (y) определяются с ошибкой, т.е. имеют случайный разброс из-за ошибок измерения и влияния неучтенных факторов. Это означает, что переменная x рассматривается как детерминированная величина, а переменная Y – как случайная, причем каждому фиксированному значению переменной x соответствует некоторое вероятностное распределение случайной величины Y.

Решение проблемы ищется в предположении, что случайная величина Y «в среднем» линейно зависит от значений переменной x, т.е. $M[Y \mid x] = ax + b$ (про это соотношение говорят, что оно описывает линейную регрессию Y на x). Тогда в качестве вероятностной модели связи между детерминированной величиной x и случайной величиной Y можно использовать уравнение $Y = ax + b + E$, интерпретировав E как случайную ошибку наблюдений с нулевым математическим ожиданием ($M[E] = 0$). (Эта модель – простейшая из линейных по параметрам регрессионных моделей.)

Построение и исследование регрессионной модели можно условно разбить на четыре этапа.

На первом этапе исследуют вопрос о наличии или отсутствии линейной связи между переменными (см. практическая работа № 10):

тем или иным способом проверяют гипотезу о значимости коэффициента корреляции. Если оснований отвергать гипотезу нет, то переходят ко второму этапу.

На втором этапе методом наименьших квадратов находят оценки параметров линейной регрессии.

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

На четвертом этапе проверяют согласованность модели с результатами наблюдений (ее адекватность).

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

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

2. Линейная регрессия. Параметры линейной регрессии. Линейная по параметрам регрессионная модель. Метод наименьших квадратов. Оценки параметров Уравнения линейной регрессии Y на x и X на y [2, с. 203 – 209, 298 – 310].

3. Ошибки наблюдения. Оценка дисперсии ошибок наблюдений. Коэффициент детерминации. Доверительные интервалы для параметров регрессии, для дисперсии ошибок наблюдений [2, с. 302 – 307; 4, с. 63 – 68].

4. Проверка значимости модели линейной регрессии [2, с. 302 – 307; 4, с. 63 – 68].

5. Проверка адекватности модели линейной регрессии [2, с. 308 – 311; 4, с. 69 – 70].

11.3. 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
imporseaborn
from sklearn.linear_model import LinearRegression
%matplotlib inline
```

Средства получения оценок линейной регрессии

Вначале надо создать экземпляр класса `LinearRegression`, который будет представлять модель регрессии: `linreg = LinearRegression()`.

Функция `linreg.fit(x, y)`

Вычисляет оценки коэффициентов регрессии b_0, b_1.

Ее параметры: `x` – двумерный массив, размера $n \times 1$ (независимая переменная); `y` – одномерный массив (вектор-строка) длины n (зависимая переменная).

Коэффициент b_0 можно получить, вызвав `linreg.intercept_`.

Коэффициент b_1 можно получить, вызвав `linreg.coef_`.

Функция `linreg.score(x, y)` вычисляет коэффициент детерминации.

11.4. Примеры для совместного обсуждения

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

В таблице приведены уровни x и y воды в реке в пунктах А и В соответственно (пункт В расположен на 50 км ниже по течению реки), замеренные в 12.00 с 1-го по 15-е апреля.

x	12,1	11,2	9,8	10,4	9,2	8,5	8,8	7,4	6,6	7,0	6,4	6,0	6,5	5,8	5,4
y	10,5	9,3	8,3	9,6	8,6	7,1	6,9	5,8	5,2	5,0	5,1	4,6	5,0	4,4	3,9

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

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

2) найдем оценки параметров β_0 и β_1, 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 β_0 и β_1.

Полученные результаты интерпретируем.

Пример 2. Считая, что зависимость между переменными имеет вид $y = \beta_0 + \beta_1 x$, найдем оценки параметров линейной регрессии по выборке:

x	1	1	2	2	3	3	2,7	2,7	4,3	4,3	4,3	5,0	5,0
y	0,5	0,1	0,5	1,2	1,2	1,7	0,9	2,2	1,1	1,7	2,5	2,0	2,2

Для выполнения задания воспользуемся инструментами языка программирования Python (вариант выполнения представлен в приложении к практической работе № 11, пример 2).

11.5. Задания для самостоятельного выполнения

В файле «Данные 11_1» приведены 30 двумерных выборок непрерывных случайных векторов. Выполните задания 1 и 2 для двумерной выборки, отобранной в соответствии с вашим вариантом.

Задание 1.

Часть 1. Осуществите статистическую обработку двумерной выборки по следующему плану.

1) Постройте диаграмму рассеивания, найдите коэффициент корреляции Пирсона; проверьте гипотезу о его значимости (выборки в файле «Данные 11_1» подобраны таким образом, что она должна подтвердиться).

2) Составьте уравнения линейной регрессии Y на x и X на y (без использования функций языка программирования Python, непосредственно находящих регрессионные прямые).

3) Нанесите графики выборочных прямых на диаграмму рассеивания.

Проанализируйте полученные результаты с точки зрения их согласованности с другими выборочными характеристиками (центром рассеивания, диаграммой рассеивания, коэффициентом корреляции).

Для контроля выполните п. 2 задания с помощью соответствующей функции языка программирования Python.

Часть 2. В предположении, что ошибки наблюдений не коррелированы и имеют нормальное распределение $N(0,\sigma)$, оцените качество аппроксимации результатов наблюдения уравнением линейной регрессии Y на x:

1) вычислите оценку s^2 для дисперсии ошибок наблюдений σ^2;

2) найдите коэффициент детерминации R^2;

3) постройте доверительные интервалы для параметров регрессии;

4) постройте доверительный интервал для дисперсии ошибок наблюдений σ^2;

5) постройте доверительные интервалы для среднего значения Y при $x = x_0$ и визуализируйте их (на рисунок с диаграммой рассеивания и регрессионными прямыми нанесите графики зависимости левой и правой границ доверительных интервалов от значения $x = x_0$).

6) проверьте статистическую значимость линейной регрессии Y на x на уровне значимости 0,05.

Все расчеты пп. 1 – 6 выполните с непосредственным использованием формул. Для контроля используйте функции библиотек python.

Задание 2. 1) 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.

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