Санкт-Петербургский государственный университет

Кафедра системного программирования

Регрессионный анализ и корреляция (живая демонстрация в Python)

Альшаеб Басель, группа 24.М71-мм

Курс «Методы статистической обработки информации»

Преподаватель: д.ф.-м.н. профессор Н.К. Кривулин

Санкт-Петербург
2025

Формат экзаменационной презентации

Цель: продемонстрировать решение задач статистического анализа на выбранном инструменте (Python: pandas, numpy, matplotlib, statsmodels).

Требование к структуре:

- краткое введение в инструмент;
- демонстрация решения задачи на примере реальных/синтетических данных;
- интерпретация результатов и выводы.

Соответствует общему плану презентации из методических указаний.

План выступления

1. Постановка задачи и данные.
2. Краткое введение в Python-инструментарий.
3. Теория: корреляция и линейная регрессия.
4. Демонстрация:
   - описательная статистика и корреляции;
   - модель OLS, сводка, проверка гипотез;
   - диагностика остатков, доверительные и предсказательные интервалы;
   - краткая связь с ANOVA (F-тест).
5. Выводы и ограничения.

Постановка задачи

Задача: изучить линейную связь между признаками и целевой переменной и построить модель для предсказания.

Пример: прогноз цены по площади и числу комнат (синтетический набор данных, размер $n = 200$).

Почему линейная регрессия и корреляция?

- простые модели;
- тесная связь с проверкой гипотез и доверительными интервалами;
- соответствуют темам курса (оценивание, гипотезы, корреляция, регрессия, ANOVA).

Инструментарий Python для демонстрации

- `pandas`: загрузка/предобработка данных, таблицы;
- `numpy`: работа с массивами, генерация данных;
- `matplotlib`: базовые графики;
- `statsmodels`: регрессия OLS, сводка, интервал оценки и предсказания.

Запуск: Jupyter/Colab/VS Code.

Установка: `pip install pandas numpy matplotlib statsmodels`

Коэффициент корреляции Пирсона

Для двух выборок (x_i) и (y_i), $i = 1, \ldots, n$, выборочная ковариация:

$$s_{XY} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}),$$

выборочные дисперсии s_X^2, s_Y^2, коэффициент корреляции:

$$r_{XY} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \in [-1, 1].$$

Интерпретация: сила *линейной* связи (не причинность). При $|r| \approx 1$ зависимость близка к линейной.

Линейная регрессия (многомерная)

Модель:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon, \quad \mathbb{E}[\varepsilon] = 0,\ \mathrm{Var}(\varepsilon) = \sigma^2.$$

Оценивание β методом наименьших квадратов (МНК):

$$\widehat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2.$$

Качество: R^2, скорректированный R^2. Проверка гипотез: $H_0 : \beta_j = 0$. Доверительные интервалы для коэффициентов и предсказаний.

Гипотезы и интервальные оценки в регрессии

- **Проверка гипотез** о параметрах: $H_0 : \beta_j = 0$ (t-стат., p-value).
- **F-тест** значимости всей модели (связь с ANOVA).
- **Доверительные интервалы** для параметров (оценки и их точность).
- **Предсказательные интервалы** для новых наблюдений (учитывают шум).

Демонстрация: генерация данных

```
import numpy as np, pandas as pd, matplotlib.pyplot as
import statsmodels.api as sm; np.random.seed(42)

n = 200
area  = np.random.uniform(30, 150, n)    # площадь
rooms = np.random.uniform(1, 5, n)       # комнаты
eps   = np.random.normal(0, 15, n)
price = 20 + 0.8*area + 10*rooms + eps   # цена

df = pd.DataFrame({"area": area, "rooms": rooms, "pric
df.head()
```

Демонстрация: описательная статистика и корреляция

```
print(df.describe(numeric_only=True))
print("\Корреляциин:\n", df.corr(numeric_only=True))

plt.figure()
plt.scatter(df["area"], df["price"])
plt.xlabel("area"); plt.ylabel("price")
plt.title("Связь␣–areaprice"); plt.show()
```

Интерпретация: знак и величина r показывают направление и силу линейной связи.

Демонстрация: модель OLS и сводка

```
X = sm.add_constant(df[["area","rooms"]])
y = df["price"]
model = sm.OLS(y, X).fit()
print(model.summary())
```

Смотрим: коэффициенты $\hat{\beta}$, их стандартные ошибки, t-статистики, p-values, R^2, F-статистику.
Вывод: если p-value для признака < 0.05, вклад статистически значим.

Демонстрация: диагностика остатков

```
resid, fitted = model.resid, model.fittedvalues

plt.figure()
plt.scatter(fitted, resid); plt.axhline(0)
plt.xlabel("Предсказания"); plt.ylabel("Остатки")
plt.title("Остатки␣vs␣предсказания"); plt.show()

sm.qqplot(resid, line="45"); plt.title("QQ–plot␣остатко
```

Интерпретация: отсутствие «веера» (гомоскедастичность), QQ-plot близок к прямой (нормальность остатков).

Демонстрация: доверительные и предсказательные интервалы

```
print("ДИ для коэффициентов:\n", model.conf_int())

new_obs = pd.DataFrame({"area":[80,120], "rooms":[2,3]
new_X   = sm.add_constant(new_obs)
pred    = model.get_prediction(new_X)
print(pred.summary_frame(alpha=0.05))   # mean_ci_lower
```

Важно: CI для среднего предсказания уже, чем PI для индивидуального наблюдения.

Связь с дисперсионным анализом (ANOVA)

F-тест из сводки `summary()` проверяет H_0: «все регрессионные коэффициенты (кроме константы) равны нулю».

Если p-value(F) < 0.05, модель в целом значима. Это соответствует идее сравнения объяснённой и остаточной дисперсий в ANOVA.

Ограничения и улучшения

- Линейность связи: при нелинейности — полиномиальные/лог-преобразования.
- Гомоскедастичность: при «веере» — робастные ошибки (Newey–West), взвешенный МНК.
- Мультиколлинеарность: регуляризация (Ridge/Lasso).
- Валидация: train/test, кросс-валидация, устойчивость результатов.

Выводы

- Корреляция и регрессия дают наглядные и интерпретируемые результаты.
- Python + statsmodels позволяют быстро пройти путь: данные $\rightarrow$ модель $\rightarrow$ гипотезы $\rightarrow$ интервалы.
- Диагностика остатков — обязательна для корректной интерпретации.

Код демонстрации можно поместить в Jupyter/Colab и приложить к сдаче.

Источники (курс)

- Конспект курса «Методы статистической обработки информации» (Кривулин Н.К.): разделы про оценивание, корреляцию, регрессию, ANOVA, интервальные оценки.
- Методические рекомендации по формату презентации и темам.