Модуль 1 · Python и анализ данных

Статистическая связь между климатическими показателями

Ph.D., старший преподаватель Мадина Турсумбаева

Кафедра метеорологии и гидрологии

Курс: Современные цифровые технологии в метеорологии

Цель лекции

Показать, как выявить, измерить и описать статистическую связь двух климатических показателей с помощью диаграммы рассеяния, коэффициента корреляции и линейной регрессии, и где проходят границы таких выводов.

Пара 1: концентрация CO_2 и глобальная температура. Пара 2: доля ВИЭ и выбросы CO_2.

Результаты обучения

После лекции студент будет:

Знать

Диаграмма рассеяния, коэффициент корреляции Пирсона, p-значение

Уравнение линейной регрессии: наклон, свободный член, R^2

Уметь

Объединить два ряда по году (merge)

Посчитать r и регрессию в pandas, SciPy и scikit-learn

Построить прогноз по времени

Понимать

r измеряет только линейную связь

Корреляция не доказывает причинность

Экстраполяция тренда ограничена

Три вопроса о связи двух величин

1. Есть ли связь?

Диаграмма рассеяния: сначала смотрим глазами

2. Насколько сильная?

Коэффициент корреляции r и его p-значение

3. Как её описать?

Уравнение регрессии: наклон и свободный член

4. Означает ли связь причину?

Статистика не отвечает на этот вопрос сама

Два примера, которые пройдут через всю лекцию

Пара 1: концентрация CO_2 (Мауна-Лоа, с 1959 г.) и глобальная аномалия температуры (GISTEMP).

Пара 2: доля ВИЭ и выбросы CO_2 по странам, в том числе в Казахстане.

Данные: объединение двух таблиц

```
temp = pd.read_csv("annual.csv")
temp = temp[temp["Source"] == "GISTEMP"][["Year", "Mean"]]
temp = temp.rename(columns={"Mean": "temp"})
co2 = pd.read_csv("co2-annmean-mlo.csv")[["Year", "Mean"]]
co2 = co2.rename(columns={"Mean": "co2"})
# соединить по ключу: году
df = pd.merge(co2, temp, on="Year")
df.tail(2)
```

Результат

	Year	co2	temp
65	2024	424,61	1,2817
66	2025	427,35	1,1850

Что в таблице

temp — аномалия температуры, °C (GISTEMP).

co2 — концентрация, ppm (Мауна-Лоа, NOAA).

Ключ объединения

merge оставляет только годы, которые есть в обеих таблицах: 67 общих лет, 1959–2025.

Шаг 1: диаграмма рассеяния

```
import matplotlib.pyplot as plt
plt.scatter(df["co2"], df["temp"], c=df["Year"])
plt.colorbar(label="Год")
plt.xlabel("CO₂, ppm"); plt.ylabel("Аномалия, °C")
```

- Точка — один год.
- Связь прямая и близкая к линейной.
- Выбросов и отдельных групп точек не видно.
- Окраска по году показывает, что точки упорядочены во времени.

Анализ связи начинается с графика, а не с числа.

Данные: NASA GISTEMP v4; NOAA GML, обсерватория Мауна-Лоа.

Шаг 2: коэффициент корреляции Пирсона

r = +1

Строго на возрастающей прямой.

r = 0

Линейной связи нет. Нелинейная связь при r = 0 возможна.

r = −1

Строго на убывающей прямой.

$$r = \Sigma\,(x - \bar{x})(y - \bar{y}) \;/\; \sqrt{\,\Sigma(x - \bar{x})^2 \cdot \Sigma(y - \bar{y})^2\,}$$

- r измеряет только линейную связь. r = 0 не означает «связи нет»: сначала смотрите на график.
- Значения r от −1 до +1; знак показывает направление связи, модуль — её силу.

Расчёт корреляции и её значимость

```
# pandas
df["co2"].corr(df["temp"])                    # 0,967
# SciPy: коэффициент и p-значение
from scipy import stats
r, p = stats.pearsonr(df["co2"], df["temp"])
r, p
```

Результат

```
(0.967, 2.1e-40)
```

r = 0,97

Очень сильная прямая связь.

p-значение

Вероятность получить такую связь, если на самом деле её нет. Обычный порог: $p < 0,05$.

Осторожно

Соседние годы зависимы: реальная значимость ниже расчётной. Во временных рядах p-значение завышено.

Шаг 3: линейная регрессия

$$y = a + b \cdot x$$

- x — предиктор (здесь CO_2); y — отклик (здесь температура).
- b — наклон: на сколько меняется y при изменении x на единицу.
- a — свободный член: значение y при x = 0.
- Остаток — вертикальное расстояние от точки до прямой.
- Метод наименьших квадратов: сумма квадратов остатков минимальна.

Что означает результат

Наклон показывает масштаб связи в единицах данных.

Уравнение описывает связь, но не объясняет механизм.

Физика, а не регрессия, объясняет связь CO_2 и температуры.

Данные: NASA GISTEMP v4; NOAA GML, обсерватория Мауна-Лоа.

Регрессия в Python: scikit-learn

```
from sklearn.linear_model import LinearRegression
X = df[["co2"]]              # таблица: двойные скобки
y = df["temp"]
model = LinearRegression()   # 1. создать
model.fit(X, y)              # 2. обучить
model.coef_[0]               # наклон b
model.intercept_             # свободный член a
model.score(X, y)            # R²
```

Результат

```
0,0108
−3,51
0,936
```

Уравнение

temp = −3,51 + 0,0108 · co2

Три шага: создать модель, обучить на данных, использовать (coef_, predict).

Как читать результат: наклон, R^2 и остатки

≈ 1,1 °C

на каждые 100 ppm CO_2 (0,0108 · 100)

R^2 = 0,94

доля объяснённой изменчивости температуры

≈ 0,09 °C

типичная ошибка модели (RMSE)

Остаток = наблюдение − модель

Большие остатки — повод разобраться: например, годы Эль-Ниньо (2016, 2024) и 1992 г. после извержения Пинатубо.

Уравнение — описание данных, а не физический закон.

Данные: NASA GISTEMP v4; NOAA GML, обсерватория Мауна-Лоа.

Корреляция не доказывает причинность

Ложная корреляция

Доля ВИЭ в Германии и глобальная температура растут вместе: r = 0,86 (1965–2024). Ветряки не нагревают планету: оба ряда растут со временем.

После удаления тренда

Берём разности соседних лет. Связь почти исчезает: r = 0,23, p = 0,07 — не значима.

```
# удаляем общий тренд: разности соседних лет
d = df[["co2", "temp"]].diff().dropna()
d["co2"].corr(d["temp"])
```

Общий тренд связывает любые растущие ряды. CO_2 и температуру связывает физика, а не сама корреляция.

Прогноз: линейная регрессия по времени

```
rows = []
for country in names:
    d = ren[ren["country"] == country]     # ren — таблица ВИЭ
    model = LinearRegression().fit(d[["Year"]], d["res"])
    pred = model.predict(pd.DataFrame({"Year": [2030, 2040]}))
    rows += [{"country": country, "Year": y, "res": v}
             for y, v in zip([2030, 2040], pred)]
forecast = pd.DataFrame(rows)
```

Страна	2030, %	2040, %	R^2
Бразилия	44,4	45,2	0,08
Германия	24,2	31,0	0,92
Китай	15,8	19,3	0,83
США	11,1	13,2	0,77
Казахстан	4,0	4,1	0,02

Проверка

Германия: прогноз на 2030 — 24,2 %; факт 2024 — 24,0 %. Прямая не описывает ускорение.

Линейный тренд: предиктор — год, отклик — доля ВИЭ. Данные: Our World in Data, тренд по 1990–2024 гг.

Экстраполяция: где заканчивается прямая

Пример из работы

Германия: линейный тренд по данным 1990–2024 даёт 31 % на 2040 год. Тот же метод на другом периоде даст другую цифру — проверьте это на своих данных.

Ограничения

Тренд описывает прошлое и не знает о будущих решениях.

Чем дальше горизонт, тем больше неопределённость.

- У прямой нет границ, а доля ВИЭ не может быть ниже 0 % и выше 100 %.
- Прогноз всегда сопровождайте интервалом или сценарием, а не одной цифрой.

Лабораторная работа: сравните прогнозы по разным периодам обучения и объясните расхождения.

Множественная регрессия: два предиктора

```
# Казахстан, 1990–2024: две таблицы → одна
kz = pd.merge(res_kz, co2_kz, on="Year")
X = kz[["Year", "res"]]            # два предиктора
y = kz["co2_per_capita"]           # т на человека в год
model = LinearRegression().fit(X, y)
model.coef_, model.intercept_, model.score(X, y)
```

Результат

```
coef_: [0,1646  −1,7382]
intercept_: −310,85
R²: 0,655
```

Уравнение модели

$CO_2 = 0{,}165 \cdot \text{Год} - 1{,}74 \cdot \text{ВИЭ} - 310{,}85$

Коэффициент

Эффект своего предиктора при неизменном втором. $R^2 = 0{,}66$.

Коэффициенты зависят от периода и выборки: сравните свои значения с лабораторной работой.

Что стоит за коэффициентом?

Коэффициент −1,74 т на 1 п. п. ВИЭ

Это причина или общий фактор?

Пример

1990-е: спад экономики — выбросы падают, а доля ВИЭ растёт сама по себе.

Вероятный общий фактор: общее потребление энергии.

Вывод

Регрессия описывает совместное изменение величин.

Причину устанавливает исследователь, опираясь на физику и теорию.

Данные: Our World in Data (energy-data, co2-data).

Типичные ошибки и их исправление

Ошибка	Что происходит	Как исправить
r = 0 понято как «связи нет»	Пропущена нелинейная связь	Сначала построить диаграмму рассеяния
Высокий R^2 принят за доказательство	Общий тренд создаёт ложную связь	Удалить тренд (diff) и проверить снова
Прогноз за пределами данных	Доля выходит за 0–100 %	Ограничить прогноз и указать горизонт
Перепутаны X и y	Модель обучена наоборот	X — предиктор, y — отклик, в fit(X, y)
Одна таблица для двух переменных	Ошибка в имени df	Разные имена: ren, co2, df

Заключение

1

График

scatter — есть ли связь

2

Сила связи

corr, pearsonr — r и p

3

Уравнение

LinearRegression — a и b

4

Остатки

y − predict — что не объяснено

Главный вывод

Связь двух величин оценивают в четыре шага: диаграмма рассеяния, коэффициент корреляции, уравнение регрессии, анализ остатков. Коэффициент r измеряет только линейную связь, корреляция не доказывает причинность, а экстраполяция тренда надёжна лишь на ограниченном горизонте.

Вопросы для самопроверки

1. Почему r = 0 не доказывает отсутствие связи?
2. Что означает наклон 0,0108 в модели temp = a + b · co2?
3. Почему p-значение во временных рядах завышено?
4. Зачем после удаления тренда брать разности соседних лет?

Семинар 3 и лабораторная работа 3

Семинар 3 · Температура и CO_2

Сравнить изменение температуры и концентрации CO_2.

Построить диаграмму рассеяния, посчитать r и регрессию.

Лабораторная работа 3 · ВИЭ и CO_2

Прогноз доли ВИЭ и выбросов CO_2 до 2030 и 2040 гг. линейной регрессией.

Вопросы: какие страны быстрее растят долю ВИЭ? Насколько точен линейный метод и какие альтернативы возможны?

Следующая лекция — Лекция 4: анализ данных метеорологических станций.

Литература

- Wilks, D. S. (2019). Statistical Methods in the Atmospheric Sciences (4th ed.). Elsevier.
- von Storch, H., & Zwiers, F. W. (1999). Statistical Analysis in Climate Research. Cambridge University Press.
- Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. JMLR, 12, 2825–2830.
- Virtanen, P., et al. (2020). SciPy 1.0. Nature Methods, 17, 261–272. https://doi.org/10.1038/s41592-019-0686-2
- GISTEMP Team (2025). GISS Surface Temperature Analysis, version 4. NASA GISS. https://data.giss.nasa.gov/gistemp/
- NOAA Global Monitoring Laboratory. Trends in Atmospheric Carbon Dioxide: Mauna Loa record. https://gml.noaa.gov/ccgg/trends/
- Ritchie, H., Rosado, P., & Roser, M. Energy; CO_2 and Greenhouse Gas Emissions. Our World in Data. https://ourworldindata.org
- VanderPlas, J. (2016). Python Data Science Handbook. O’Reilly Media. https://jakevdp.github.io/PythonDataScienceHandbook/