**Statsmodels**

**Что такое statsmodels и чем отличается от scikit-learn**

statsmodels и scikit-learn — это две библиотеки Python для статистического анализа и машинного обучения соответственно, но они служат разным целям и философия у них тоже разная.

**📊 Что такое statsmodels?**

**statsmodels** — библиотека для проведения **традиционного статистического анализа**, особенно в эконометрике и социальных науках. Она предоставляет:

* **Обыкновенные наименьшие квадраты (OLS)** и другие регрессионные модели
* **Обобщённые линейные модели (GLM)**
* **Временные ряды**: ARIMA, SARIMAX и др.
* **Статистические тесты**: t-тесты, тесты на нормальность, автокорреляцию и т.д.
* **Оценка доверительных интервалов и p-value**
* **Подробная статистика и отчёты**, как в R

import statsmodels.api as sm

X = sm.add\_constant(data[['feature1', 'feature2']])

y = data['target']

model = sm.OLS(y, X).fit()

print(model.summary()) # Таблица с коэффициентами, p-value, R² и т.д.

## 🤖 Что такое scikit-learn?

**scikit-learn (sklearn)** — библиотека для **машинного обучения**, особенно в инженерных и прикладных задачах. Она охватывает:

* Классификацию, регрессию, кластеризацию
* Обучение с учителем и без
* Пайплайны, кросс-валидацию, GridSearch
* PCA и другие методы снижения размерности
* Интерфейсы совместимы со NumPy/Pandas

### Пример:

from sklearn.linear\_model import LinearRegression

model = LinearRegression()

model.fit(X, y)

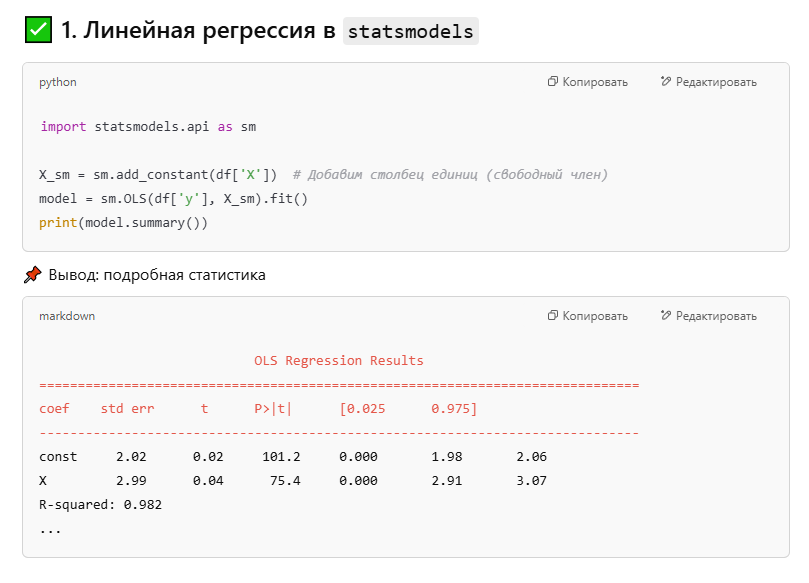
print(model.coef\_, model.intercept\_)

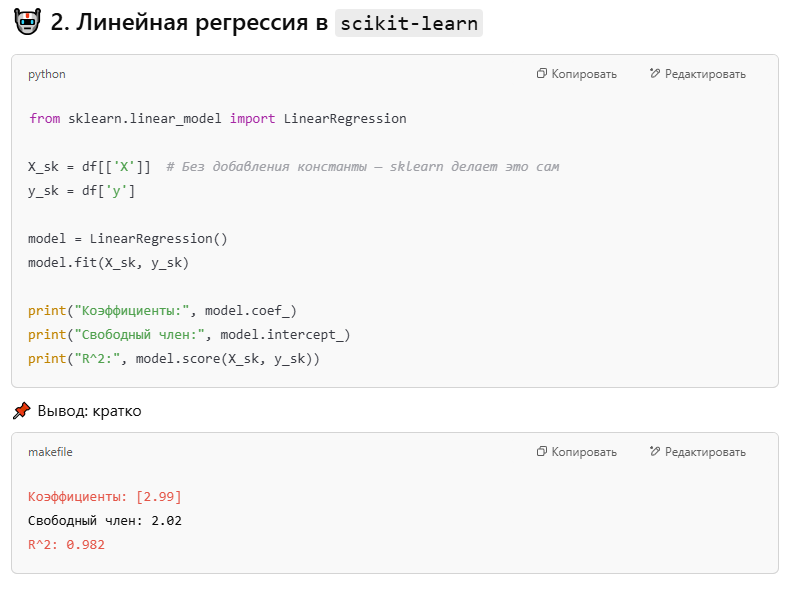


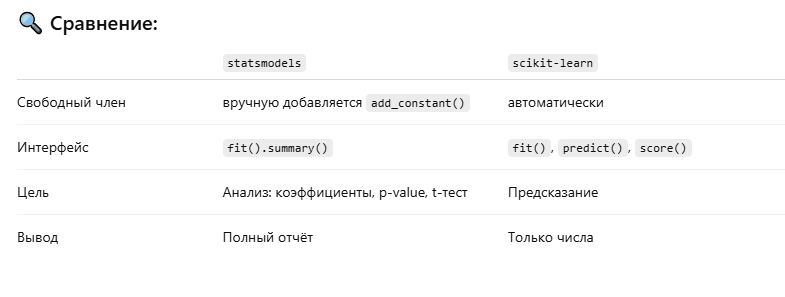
**🧠 Когда использовать что?**

* **Если вы исследуете данные, делаете отчёт, проверяете гипотезы, работаете с эконометрикой** → используйте statsmodels.
* **Если вы строите предсказательные модели, пайплайны или участвуете в ML-проектах** → используйте scikit-learn.



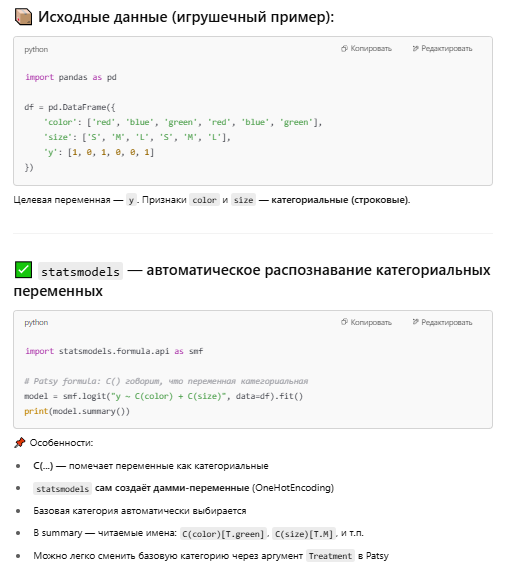


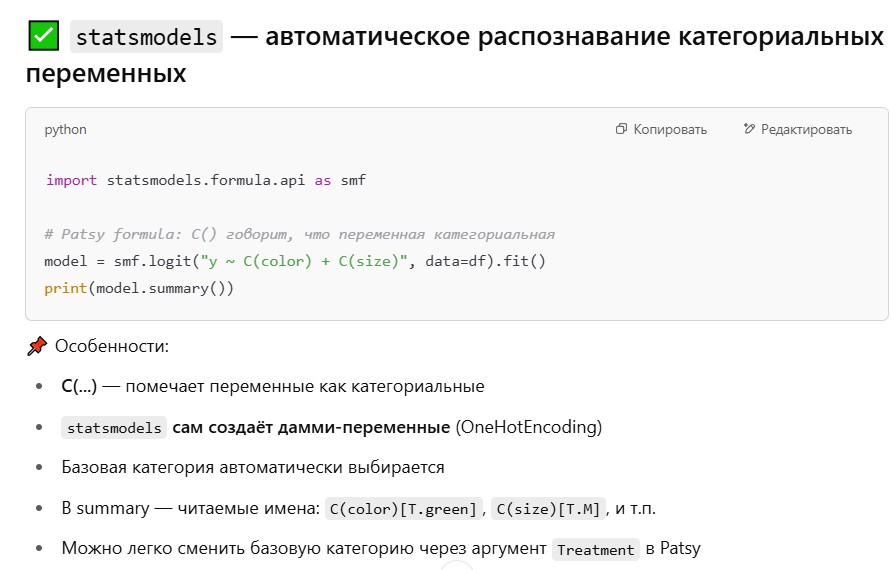


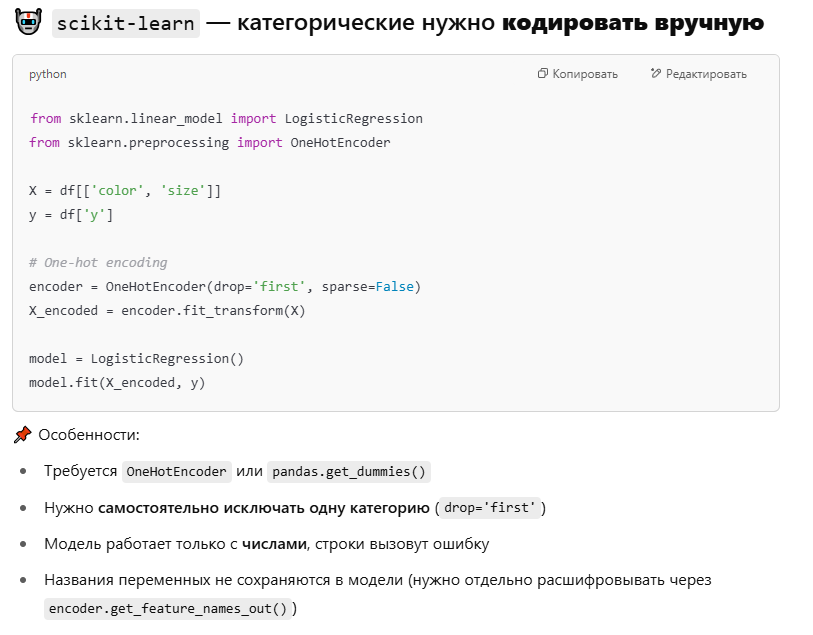


**Сравнение работы с категориальными переменными в обеих библиотеках**

**Работа с категориальными переменными** (например, пол, цвет, страна и т.п.) — ключевая часть анализа данных. statsmodels и scikit-learn подходят к ней по-разному:

****

****

****

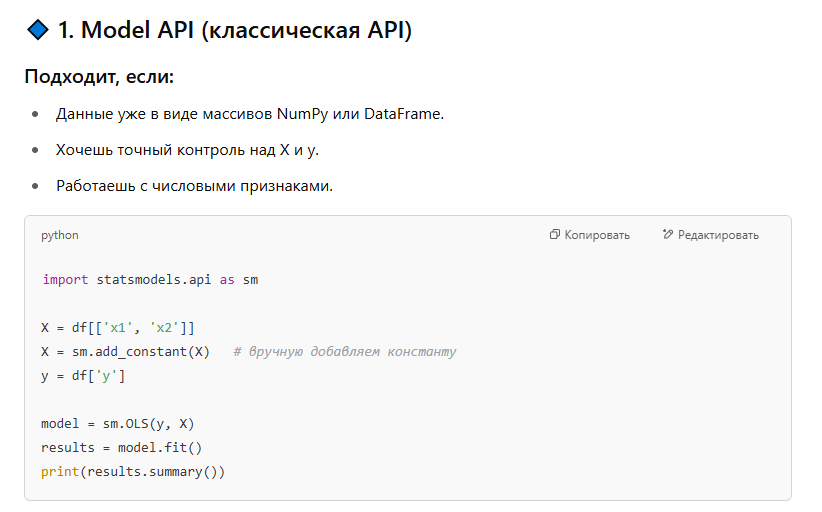
****

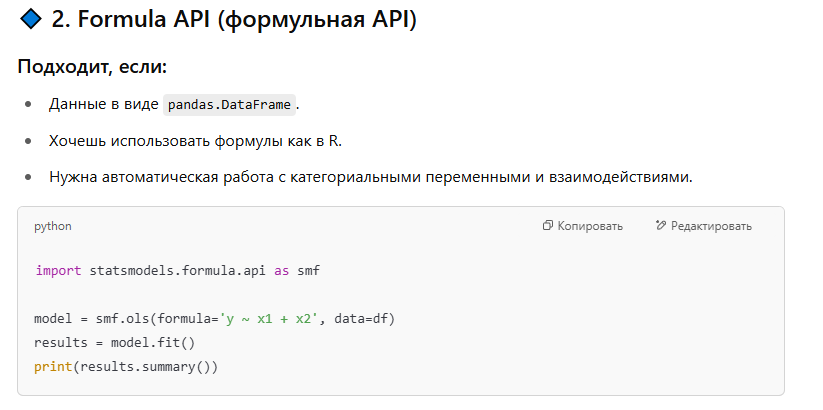
**🧠 Вывод**

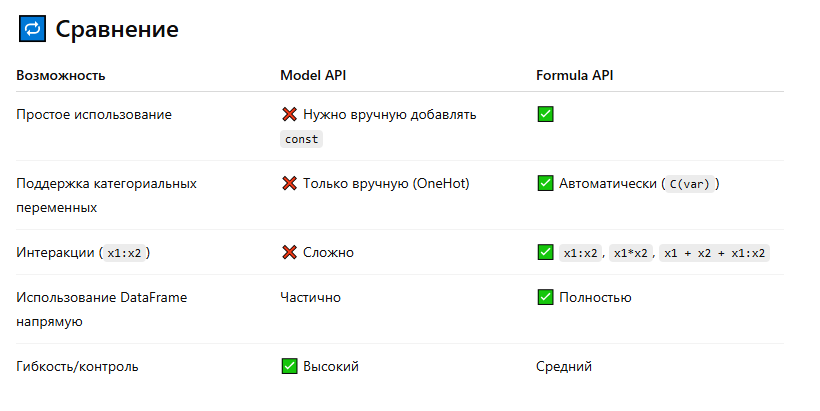
* statsmodels лучше подходит для **анализа категориальных переменных**, особенно если тебе нужны p-value, доверительные интервалы и интерпретируемые формулы.
* scikit-learn требует **ручной подготовки признаков**, зато хорошо масштабируется и интегрируется в пайплайны машинного обучения.

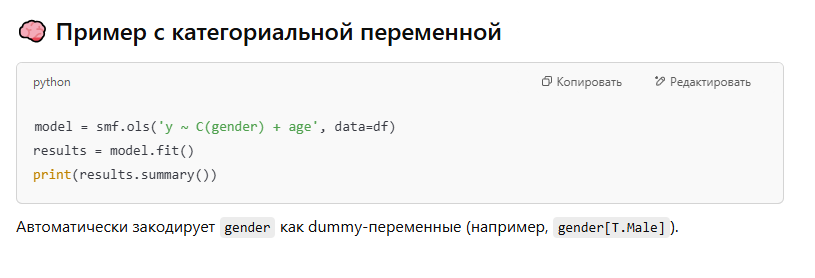
**API: формулы (formula API) и model API**

В statsmodels есть **две основные парадигмы создания моделей**: **formula API** и **model API**. Обе ведут к одной и той же статистической модели, но различаются **в способе задания входных данных** и **в удобстве работы с категориальными переменными и взаимодействиями**.









**Обзор модулей: statsmodels.api, statsmodels.formula.api**

**🔹 statsmodels.api — Model API (объектно-ориентированный подход)**

Это **основной модуль**, через который доступна большая часть функционала statsmodels.

**✅ Особенности:**

* Ориентирован на **явную передачу данных (матрицы X и y)**.
* Позволяет точнее контролировать предобработку.
* Подходит для **программного построения моделей**, когда данные не всегда известны заранее.

****

**🧰 Что включает:**

* OLS, GLS, GLM, Logit, Poisson, RLM, WLS, MixedLM и др.
* Доступ к диагностике: graphics, distributions, datasets.

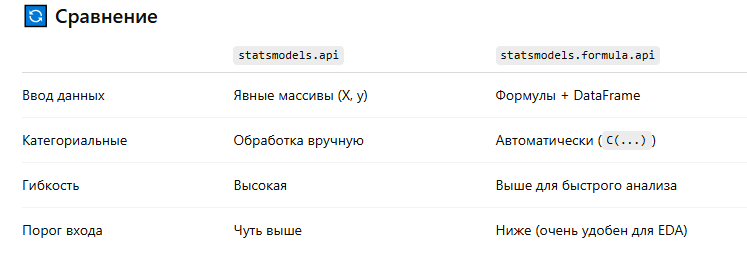
### 🔹 statsmodels.formula.api — **Formula API (формульный подход, как в R)**

Этот модуль позволяет задавать модели через **формулы** в стиле R, используя patsy под капотом.

#### ✅ Особенности:

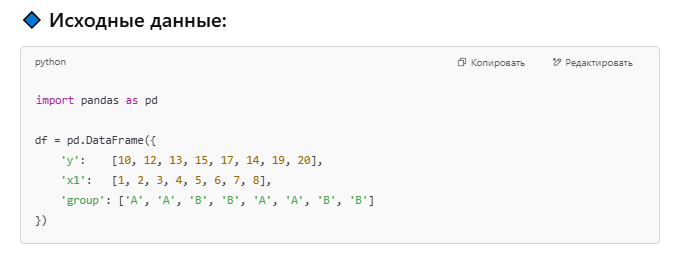
* Работает с **pandas.DataFrame**.
* Очень удобно для **быстрого моделирования с категориальными переменными** и трансформациями.
* Автоматически добавляет константу.

****

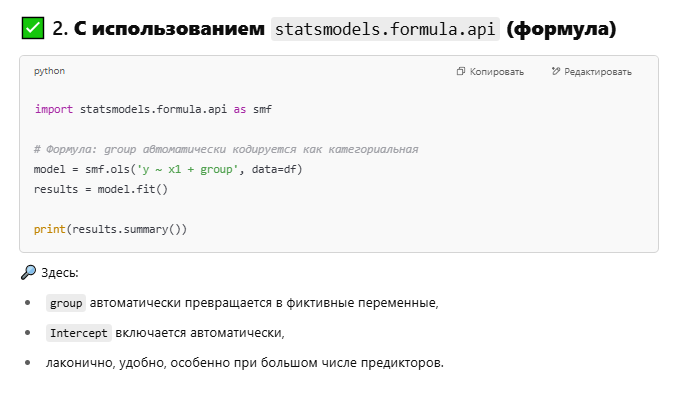
****

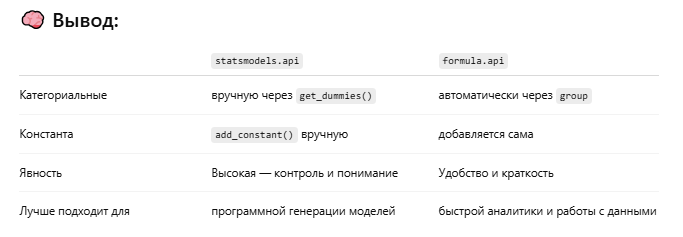
**Как одна и та же модель записывается в обоих стилях на одном примере**

Рассмотрим один и тот же пример — **линейную регрессию**:  
целевая переменная y, предикторы — числовой x1 и категориальный group.

****

****

****

****