

# ЛЕКЦИЯ 7. Качественные переменные в регрессии: как добавить "да" и "нет" в уравнение

## Введение

На предыдущих лекциях вы не только научились строить регрессионные модели, но и начали смотреть на них критично.

Вы уже умеете:

подбирать количественные переменные,

собирать таблицы,

применять регуляризацию,

проверять допущения,

и даже **замечать, когда переменные "дублируют" друг друга**, вызывая мультиколлинеарность.

Но все эти переменные до сих пор имели одну общую черту:

👉 они были **числовыми**. То есть вы могли напрямую измерять:

**Сколько часов сна? Какой уровень стресса? Сколько баллов по шкале?**

Однако в реальности вы часто сталкиваетесь с другим типом информации — **качественной**:

Пол: мужчина или женщина


Курит или не курит

Работает или не работает

Категория образования или профессии

Такие переменные — важны, но напрямую их **нельзя вставить в уравнение**.

Нельзя написать: "женщина \* 3.5" — модель этого не поймёт.

 Именно здесь начинается новая важная глава в вашем аналитическом мышлении:

**как превратить качественные признаки в то, с чем может работать регрессия, и как не потерять при этом смысл.**

Сегодня вы узнаете:

что такое фиктивные переменные (dummy variables),

как они кодируются,

как интерпретируются,

и как сделать так, чтобы "женщина", "курит" или "бакалавр" стали частью вашей модели — **корректно, логично и понятно.**

## Почему это важно?

Почти каждая реальная гипотеза содержит **не только числовые, но и логические переменные:**

"У женщин уровень стресса при одинаковой нагрузке выше, чем у мужчин"

"Работающие студенты спят меньше, чем неработающие"

"Курение влияет на уровень энергии независимо от физической активности"

Чтобы проверить такие гипотезы, нужно добавить в модель **качественные переменные**, и для этого используется приём, который называется **фиктивные переменные** (dummy variables).

## Что такое фиктивная переменная?

Это способ **перевести категорию в 0 и 1:**

Женщина = 1, Мужчина = 0

Работает = 1, Не работает = 0

Курит = 1, Не курит = 0

Теперь это можно вставить в уравнение как обычную переменную.

## Пример гипотезы

**Гипотеза:**

Женщины спят меньше мужчин при одинаковом уровне нагрузки.

### Таблица

Респондент	Стресс ( $x_1$ )	Пол (женщина=1)	Время сна ( $y$ )
1	3	1	6.5
2	5	0	7.0
3	2	1	6.0
4	4	0	7.5
5	6	1	6.2

### Формула модели

$$y = b_0 + b_1 * \text{стресс} + b_2 * \text{женщина}$$

Здесь:

$b_1$  — как стресс влияет на сон

$b_2$  — разница во сне **между женщинами и мужчинами**

$b_0$  — базовое значение сна **у мужчин**, при стрессе = 0

### Как интерпретировать $b_2$ ?

Если  $b_2 = -0.6$ , это означает:

При одинаковом уровне стресса женщины спят в среднем на **0.6 часа меньше**, чем мужчины.

Если  $b_2 = 0.4$ :

Женщины спят **на 0.4 часа больше**, при одинаковом уровне стресса.

### Как построить модель в Excel / Google Sheets

Google Sheets:

```
=LINEST(D2:D6, A2:B6, TRUE, TRUE)
```

Excel (русский язык):

```
=ЛИНЕЙН(D2:D6; A2:B6; ИСТИНА; ИСТИНА)
```

Где:

A2:A6 — стресс

B2:B6 — пол (женщина = 1, мужчина = 0)

D2:D6 — время сна

## Что делать с переменными с 3 и более категориями?

Пример: образование — бакалавр / магистр / PhD

Выбираете **базовую категорию** (например, бакалавр), а остальные кодируете:

Образование	Магистр	PhD
Бакалавр	0	0
Магистр	1	0
Доктор наук (PhD)	0	1

**Нельзя добавлять все 3 столбца** — это создаст **идеальную корреляцию** (см. мультиколлинеарность).

Нужно всегда **исключать одну категорию** — она становится **“базовой”**.

### ⚠ Частые ошибки

- ❌ Вставить категорию как текст (“мужчина”, “женщина”)
- ❌ Добавить все 3 фиктивные переменные вместо  $k-1$
- ❌ Не указать, какая категория является базовой
- ❌ Интерпретировать  $b_2$  без понимания, **что означает 0, а что 1**

## Использование ИИ

Инструмент	Для чего
ChatGPT	Поможет интерпретировать фиктивную переменную
Excel Copilot	Автоанализ категориальных данных
Notion AI	Поможет описать, что означает каждый коэффициент

### 🚫 Запрещено

Не переводить категорию в числа

Использовать текстовые значения напрямую

Игнорировать смысл "базовой" категории

Упрощать: "ну просто добавлю, чтобы было"

## **Вывод**

Качественные переменные — это мост между реальной жизнью и цифрами.

Если вы умеете их правильно кодировать, вы расширяете аналитические возможности модели и делаете её ближе к реальному человеку.