

# Лекция 5 — Множественная линейная регрессия

## Введение

В предыдущих лекциях вы шаг за шагом создавали свою первую аналитическую модель:

1. **Вы формулировали гипотезу**, как исследователь — не как статистик.
2. **Переводили её в цифры**: выбирали переменные, оцифровывали ответы, строили таблицу.
3. **Применяли метод наименьших квадратов**, чтобы подогнать линию, объясняющую зависимость.
4. А затем — **F-тест**, чтобы спросить себя:

*“А имеет ли смысл моя модель вообще?”*



F-тест — это первый шаг в **критическом мышлении аналитика**:  
“Не просто что получилось, а **можно ли этому верить?**”

## Что дальше?

Вы уже умеете строить регрессию по **одной переменной**.

Но мир гораздо сложнее.

В большинстве жизненных и исследовательских ситуаций:

на поведение человека влияет не один фактор;

на результат влияет не только то, что вы **намеренно изучаете**, но и то, что **скрыто в контексте**;

простая модель оказывается **недостаточной**.

Вот тут появляется следующий уровень мышления: **Множественная линейная регрессия**.

## Почему одной переменной мало?

*Представьте:*

Вы изучаете, влияет ли экран перед сном на качество сна.

Получаете отрицательную связь: экран  $\downarrow \rightarrow$  сон  $\downarrow$ .

*Но...*

Что если **стресс** сильнее влияет на сон, чем экран?

А экран — просто **сопровождает стресс**, но сам по себе не так важен?

Если вы **не включите стресс в модель**, то приписываете экрану **влияние, которого у него нет**.

**Это и есть ошибка упущенной переменной.**

Она приводит к ложным выводам.

Вы "ловите" зависимость, которой на самом деле нет.

**Чтобы этого избежать, вам нужно:**

включать все важные переменные,

оценивать их влияние **одновременно**,

и проверять, **какой вклад** вносит каждая.

## К чему мы сейчас переходим

*В этой лекции вы научитесь строить регрессию с **двумя и более объясняющими переменными**;*

читать формулу, где у каждого  $x$  — свой коэффициент влияния;

интерпретировать **влияние каждого признака**, с учётом других;

находить, **какие переменные действительно работают**, а какие — нет;

готовить таблицы и рассчитывать модель в **Google Sheets и Excel**.

## Как связать это с F-тестом?

*F-тест говорит:*

"Да, модель в целом значима."

Теперь вы переходите к следующему вопросу:

"А что именно в ней работает?"

Множественная регрессия — это уже **исследование системы**, а не одной зависимости.

## Что такое множественная регрессия?

Это расширение обычной (простой) линейной регрессии, в которой на результат у влияет **не одна**, а **несколько переменных** одновременно.

 Формула:

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + ... + b_n * x_n$$

Где:

y — зависимая переменная (то, что вы хотите объяснить),

$x_1, x_2, x_3, ... x_n$  — независимые переменные (факторы),

$b_0$  — свободный коэффициент (уровень y при нулевых x),

$b_1, b_2, b_3...$  — влияние каждого x на y.

## Пример гипотезы

Качество сна студентов зависит от:

количества времени у экрана перед сном ( $x_1$ ),

уровня стресса ( $x_2$ ),

времени отхода ко сну ( $x_3$ ).

## Пример таблицы данных

Респондент	Экран ( $x_1$ )	Стресс ( $x_2$ )	Время сна ( $x_3$ )	Качество сна (y)
1	2	3	1.5	6
2	4	4	23.5	7
3	0	1	0.25	8
4	8	7	2.0	3
5	5	5	1.0	5

**Время сна переведено в десятичный формат:**

– 01:30 → 1.5

– 23:30 → 23.5

– 00:15 → 0.25

## Как построить множественную регрессию в Excel / Google Sheets

**Google Sheets (английский язык):**

```
=LINEST(E2:E6, A2:C6, TRUE, TRUE)
```

**Excel (русский язык):**

```
=ЛИНЕЙН(E2:E6; A2:C6; ИСТИНА; ИСТИНА)
```

Где:

A2:C6 — ваши независимые переменные  $x_1, x_2, x_3$ ,

E2:E6 — зависимая переменная  $y$  (качество сна)

### Интерпретация коэффициентов

Допустим, результат:

$b_1 = -0.6$  (экран)

$b_2 = -0.4$  (стресс)

$b_3 = 0.3$  (время сна)

$b_0 = 7.2$  (свободный коэффициент)

$R^2 = 0.87$

### Что это значит:

$b_1 = -0.6$  → экран ухудшает сон: каждая единица увеличения экрана снижает оценку сна на 0.6 балла.

$b_2 = -0.4$  → стресс также ухудшает сон, но слабее.

$b_3 = 0.3$  → чем позже ложится человек (в этой шкале), тем лучше сон? 🤔  
Возможно, нужна другая шкала!

$b_0 = 7.2$  → качество сна при нулевом экране, нулевом стрессе и "нулевом времени"

$R^2 = 0.87 \rightarrow$  модель объясняет 87% колебаний  $y$  — это сильно.

## Как читать модель

Элемент модели	Интерпретация
$b > 0$	Переменная повышает результат
$b < 0$	Переменная снижает результат
$b \approx 0$	Не влияет или шум
$R^2$ близок к 1	Модель сильная
$R^2 < 0.5$	Модель объясняет мало

## Шкалы: как кодировать переменные

Чтобы модель была корректной, **все переменные должны быть числовыми и логичными.**

Примеры:

### Стресс:

Ответ	$\rightarrow$ Балл
Никогда	10
Иногда	6
Часто	3
Всегда	0

### Время отхода ко сну:

23:00  $\rightarrow$  23.0

00:30  $\rightarrow$  0.5

02:00  $\rightarrow$  2.0

### ⚠ Частые ошибки

- ❌ Переменные закодированы нелогично (например, "больше баллов" = "хуже")
- ❌ Интерпретация коэффициентов без учета масштаба
- ❌ Переменные "дублируют" друг друга  $\rightarrow$  мультиколлинеарность

✗ Нет F-теста → модель может быть незначимой

## Как понять важность переменной?

Сравнивайте **модули коэффициентов**

Смотрите, **меняется ли  $R^2$** , если исключить переменную

Постройте модель с и без переменной — сравните результат

## Использование ИИ

Можно применять:

Инструмент	Для чего
Excel Copilot	Подскажет значимые переменные, объяснит модель
Notion AI	Поможет сформулировать выводы
ChatGPT	Интерпретация коэффициентов и $R^2$
Sheets Explore	Быстрый анализ таблицы и автоматическая регрессия

## 🚫 Категорически запрещено:

Использовать множественную модель “по шаблону” без понимания

Не описывать, что означает каждый  $x$


Включать переменные “на всякий случай”

Игнорировать F-статистику, если она низкая

## Вывод

Множественная линейная регрессия — это **переход от простого анализа к системному мышлению**.

Вы больше не ищете “одну причину”, вы строите модель **множества взаимосвязей**, оцениваете их вклад и силу.

 **Это уже уровень настоящего исследователя, продукта, аналитика.**