ЛЕКЦИЯ 6. Мультиколлинеарность: когда переменные спорят между собой

Введение

В предыдущей лекции вы узнали, как построить множественную регрессию — модель, которая учитывает несколько переменных, влияющих на результат. Вы научились:

формировать сложную гипотезу,

кодировать переменные,

анализировать вклад каждой из них,

интерпретировать коэффициенты.

Но чем больше переменных вы добавляете, тем выше риск, что они **мешают друг другу**.

Именно это явление называется **мультиколлинеарностью**.

Представьте:

Вы хотите предсказать качество сна.

Включаете в модель: время у экрана, стресс, тревожность, внутреннее беспокойство.

Проблема?

Тревожность и внутреннее беспокойство почти одно и то же.

Модель начинает “путаться”:

Не знает, кому “отдать влияние” на результат.

Коэффициенты становятся нестабильными.

Значимость переменных может обнуляться.

Что такое мультиколлинеарность?

Это ситуация, когда две или более объясняющих переменных слишком сильно коррелируют между собой.

Визуально:

$x_2 \approx x_3$ → модель не может различить, кто из них важнее

Почему это плохо?

Последствие	Объяснение
Коэффициенты становятся нестабильными	Чуть меняются данные — сильно меняется результат
R^2 может быть высоким, но переменные незначимы	Кажется, что модель хорошая, но ни одна переменная не “работает”
Интуитивные переменные теряют силу	Логичные признаки получают “0” в коэффициенте

Признаки мультиколлинеарности

1. **Высокая корреляция между** x **переменными**

(например, > 0.8)

2. **Одна из переменных имеет “странно низкий” коэффициент**

(или даже противоположный знак)

3. **Высокий R^2, но все** p-value **переменных > 0.05**

(т.е. модель “вроде бы хорошая”, но ничего не значимо)

Пример

Вы строите модель:

$$\text{Качество сна} = b_0 + b_1 * \text{стресс} + b_2 * \text{тревожность}$$

Но стресс и тревожность — почти одно и то же.

Получаете:

$R^2 = 0.82$

b_1 = 0.2 (p = 0.41)

b_2 = -0.3 (p = 0.37)

> **Итог:** модель сильная, а переменные “не работают”. Это и есть мультиколлинеарность.

Как найти мультиколлинеарность

1. **Постройте корреляционную матрицу X-переменных**

(в Google Sheets: =CORREL(x1_range, x2_range))

2. Если корреляция между переменными > 0.7–0.8 — возможна проблема
3. Визуально: если два столбца почти “копируют” друг друга — опасный сигнал

Что делать, если мультиколлинеарность есть?

Метод	Что делать
Удалить одну из переменных	Оставьте более информативную или понятную
Объединить переменные	Например: стресс + тревожность → индекс выгорания
Провести факторный анализ	(расширенный подход, можно с поддержкой преподавателя)
Использовать регуляризацию	Об этом — в лекциях про Ridge и Lasso (впереди!)

Excel-инструменты

1. Вычислите =CORREL(...) между парами переменных
2. Проверьте матрицу на сильную корреляцию
3. Постройте множественную модель **с разными наборами переменных**, и сравните:

Как меняется R^2

Как меняются коэффициенты

Что происходит с p-value

Что ИИ может (и должен) делать

Инструмент	Как использовать
ChatGPT	Объяснит, почему коэффициенты “вдруг обнулились”
Excel Copilot	Покажет визуализацию зависимости между переменными
Notion AI	Поможет описать структуру взаимосвязей

🚫 Что запрещено

Игнорировать мультиколлинеарность “потому что R^2 высокий”

Оставлять одинаковые переменные, чтобы “сдача была полной”

Не объяснять, почему убрали переменную

Ставить “тревожность”, “выгорание”, “эмоциональную усталость” как отдельные признаки без пояснения

Вывод

Мультиколлинеарность — это **невидимая угроза** внутри вашей модели.

Она разрушает интерпретацию и делает модель внешне красивой, но внутренне недостоверной.