

Лекция 8. Фиктивные переменные

Введение. Как измерить неизмеримое?

До сих пор мы работали с количественными переменными (зарплата, образование, цена). Но как включить в модель такие качественные признаки, как **пол, отрасль, регион, наличие высшего образования, квартальный эффект**?

Фиктивные переменные - это искусственно созданные переменные, которые принимают значения 0 или 1 и используются для кодирования принадлежности наблюдения к определенной категориальной группе.

Аналогия: Представьте, что вы изучаете эффективность двух лекарств (А и Б). Вы не можете "измерить" лекарство в цифрах. Но вы можете создать две фиктивные переменные:

- $D_A = 1$, если пациент получал лекарство А, и 0 в противном случае.
- $D_B = 1$, если пациент получал лекарство Б, и 0 в противном случае.

Теперь вы можете количественно оценить эффект каждого лекарства по сравнению с контрольной группой.

1. Фиктивные переменные для двух категорий

Самый простой случай — бинарный признак (например, пол).

Модель:

$$Y_i = \beta_0 + \beta_1 X_i + \delta D_i + \varepsilon_i$$

где $D_i = 1$ для мужчин, $D_i = 0$ для женщин.

Интерпретация:

- Для женщин ($D=0$): $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
- Для мужчин ($D=1$): $Y_i = (\beta_0 + \delta) + \beta_1 X_i + \varepsilon_i$

- β_0 — среднее значение Y для женщин при $X=0$.
- δ — разница в среднем значении Y между мужчинами и женщинами при одном и том же уровне X .

2. Фиктивные переменные для нескольких категорий (ловушка фиктивных переменных)

Когда категорий больше двух (например, Регион: Север, Юг, Восток), мы создаем несколько фиктивных переменных.

Правило: для категориальной переменной с m категориями нужно ввести $m-1$ фиктивную переменную. Если ввести m переменных, возникнет **совершенная мультиколлинеарность** (ловушка фиктивных переменных).

Пример для 3х регионов:

- $D_{Юг} = 1$ если регион Юг, и 0 иначе.
- $D_{Восток} = 1$ если регион Восток, и 0 иначе.
- **Базовая (контрольная) категория:** Север (когда обе фиктивные переменные равны 0).

Модель:

$$Y_i = \beta_0 + \beta_1 X_i + \delta^1 D_{Юг_i} + \delta^2 D_{Восток_i} + \varepsilon_i$$

Интерпретация:

- β_0 — средний Y для Севера (базовая категория).
- δ^1 — разница в среднем Y между Югом и Севером (при одинаковом X).
- δ^2 — разница в среднем Y между Востоком и Севером (при одинаковом X).

3. Взаимодействие фиктивных переменных с количественными переменными

До сих пор мы предполагали, что наклон одинаков для всех групп. Но что, если влияние X на Y различается для разных групп? (Например, "цена" года образования разная для мужчин и женщин).

Для этого в модель вводятся **переменные взаимодействия**.

Модель с взаимодействием:

$$Y_i = \beta_0 + \beta_1 X_i + \delta D_i + \gamma(X_i * D_i) + \varepsilon_i$$

Интерпретация:

- Для группы **$D=0$** : $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
 - Наклон: β_1
- Для группы **$D=1$** : $Y_i = (\beta_0 + \delta) + (\beta_1 + \gamma)X_i + \varepsilon_i$
 - Наклон: $\beta_1 + \gamma$
- γ показывает, **насколько отличается наклон** для группы $D=1$ от наклона для базовой группы.

4. Тестирование гипотез с фиктивными переменными

4.1. Тест на различие констант (Chow Test для структурного сдвига)

- **Гипотеза:** $H_0: \delta = 0$ (нет различий в среднем уровне Y между группами).
- **Тест:** Обычный t-тест для коэффициента δ .
- **Вывод:** Если $p\text{-value} < \alpha$, различия в константах статистически значимы.

4.2. Тест на различие наклонов

- **Гипотеза:** $H_0: \gamma = 0$ (наклоны регрессии одинаковы для групп).
- **Тест:** Обычный t-тест для коэффициента γ .

- **Вывод:** Если $p\text{-value} < \alpha$, различия в наклонах статистически значимы.
-

4.3. Тест на полное различие моделей

- **Гипотеза:** Все коэффициенты, связанные с фиктивными переменными, равны нулю.
- **Тест:** F-тест, сравнивающий исходную модель с моделью, включающей все фиктивные переменные и взаимодействия.

5. Сезонные фиктивные переменные

Частый случай применения - учет сезонности в данных временных рядов.

Пример: Ежеквартальные данные.

- $Q1 = 1$ если 1-й квартал, и 0 иначе.
- $Q2 = 1$ если 2-й квартал, и 0 иначе.
- $Q3 = 1$ если 3-й квартал, и 0 иначе.
- **Базовая категория:** 4-й квартал.

Модель:

$$Y_t = \beta_0 + \beta_1 X_t + \delta_1 Q1_t + \delta_2 Q2_t + \delta_3 Q3_t + \varepsilon_t$$

- δ_1 показывает, на сколько в среднем Y в 1-м квартале отличается от Y в 4-м квартале.

6. Практический пример: Зарплата, образование и пол

Исходная модель (без взаимодействия):

$$\text{Зарплата} = 20 + 4 * \text{Образование} + 5 * \text{Пол_Мужчина}$$

- Для женщин: $\text{Зарплата} = 20 + 4 * \text{Образование}$
- Для мужчин: $\text{Зарплата} = 25 + 4 * \text{Образование}$

- **Вывод:** При одинаковом образовании мужчины получают на 5 тыс. руб. больше.

•

Модель с взаимодействием:

$$\text{Зарплата} = 18 + 5 * \text{Образование} + 8 * \text{Пол_Мужчина} - 1 * (\text{Образование} * \text{Пол_Мужчина})$$

- Для женщин: Зарплата = 18 + 5*Образование
- Для мужчин: Зарплата = 26 + 4*Образование
- **Вывод:**
 - У женщин "цена" года образования = 5.
 - У мужчин "цена" года образования = 4.
 - Мужчины имеют более высокую стартовую зарплату (константа), но отдача от образования у них ниже.

Резюме

1. **Фиктивные переменные** позволяют включать в регрессию качественные факторы.
2. Для **m категорий** создаем m-1 переменную, чтобы избежать ловушки.
3. Коэффициент при фиктивной переменной показывает **сдвиг константы** относительно базовой категории.
4. **Переменные взаимодействия** позволяют моделировать различное влияние количественных переменных на разные группы.
5. С помощью **t-тестов** и **F-тестов** можно проверить статистическую значимость различий между группами.
6. **Сезонные фиктивные переменные** — мощный инструмент для учета сезонности.
- 7.

На следующей лекции: Мы перейдем к основам анализа **временных рядов**, где изучим понятия стационарности и единичных корней.

Вопросы для самопроверки:

1. Почему нельзя включить фиктивные переменные для всех категорий одновременно?
2. Как бы вы закодировали фиктивные переменные для переменной "Уровень образования" (Среднее, Бакалавр, Магистр)?
3. В модели с взаимодействием коэффициент при фиктивной переменной оказался незначим, а коэффициент при взаимодействии - значим. Как это интерпретировать?