Мультиколлинеарность

Линейная регрессия

Novosibirsk State University
*THE REAL SCIENCE

2GIS

Мультиколлинеарность — наличие значимой линейной взаимосвязи между независимыми переменными (факторами) в регрессионной модели.

Мультиколлинеарность:

- **Строгая (полная).** Функциональная взаимосвязь между факторами.
- **Нестрогая (частичная).** Сильная корреляционная связь между факторами.

Полная мультиколлинеарность

В модели $Y = const + b_1X_1 + b_2X_2 + \ldots + b_mX_m + \varepsilon$

существуют взаимосвязи вида

$X_1 = const + a_1X_2$

Почему это плохо?

1. Нет единственной оценки МНК (уравнения с совсем разными коэффициентами могут давать результат одного качества).
2. Невозможно отделить эффекты каждого фактора в отдельности.

Причина:

Ошибка исследователя (на этапе отбора признаков что-то пошло не так). Например, распространенная ошибка: включение лишней фиктивной переменной.

Частичная мультиколлинеарность

Какие факторы бывают связаны?

1. Одно и то же явление измерено немного по-разному.
2. Измерены разные вещи, но они естественно и логично связаны между собой.

Частичная мультиколлинеарность — это плохо?

Можно ничего не делать.

В отличие полной (строгой), частичная мультиколлинеарность не нарушает предположений о регрессионной модели.

Но может стать проблемой, когда коэффициенты корреляции большие (>0,8-0,9).

Проблемы, связанные с частичной мультиколлинеарностью

- Увеличивает стандартные ошибки, занижает t-статистики, соответственно, снижает устойчивость модели и точность прогноза.
- Остаётся проблема отделения эффектов каждого отдельного фактора (когда они объясняют не только отклик, но и друг друга).
- Бывают ситуации, когда взаимосвязанность факторов «скрывает» влияние каждого из них на отклик.

Как обнаружить мультиколлинеарность: некоторые симптомы

- Высокий R^2 при низких t (или большом количестве незначимых коэффициентов).
- Странные (логически необъяснимые) знаки регрессионных коэффициентов.
- Небольшие изменения в данных существенно меняют модель (знаки и значения оценок).
- Коэффициент VIF (variance inflation factor) рассчитывается для каждого фактора и показывает степень увеличения дисперсии за счёт коррелированности каждого фактора с остальными.
  VIF = 1. Значит, проблемы нет.
  VIF > 10. Мультиколлинеарность.

Как бороться с мультиколлинеарностью?

- Тщательно анализировать данные «на входе»: исследовать распределения и взаимосвязи между факторами (и откликом).

Если мультиколлинеарность пробралась в модель, то:

- Исключить одну из переменных.
- Преобразовать коррелированные переменные.
- Ничего не делать.