

Типы данных

Линейная регрессия

Технические ограничения на типы переменных при построении линейной регрессии

Можно использовать:

- Метрические (количественные) переменные.
- Порядковые (ранговые) переменные.

Использование данных: ограничения

Порядковые переменные:

- Используются в линейной регрессии только в качестве фактора (отклик может быть только **количественным**).
- Количество рангов должно быть достаточно велико.
- Расчёт «среднего» должен иметь содержательный смысл.

Бинарные (дихотомические) данные также пригодны для использования в модели.

Интервальные данные:

- Не должны содержать выбросов и других аномалий.

Номинальные данные:

- В чистом виде не используются.
- Можно включать, если перекодировать в набор фиктивных переменных.
- Если признак принимает k значений (есть k категорий), то создается $k - 1$ переменных, принимающих значения 0 – 1.

Прежде, чем строить модель, убедитесь, что:

1. Все интервальные данные очищены от выбросов и неопределённых значений.
2. Все факторы связаны с откликом (и слабо связаны между собой).
3. Все признаки нужного типа, если нет — что они преобразованы.