Мультиколлинеарность

Эконометрика. Openedu. Неделя 6

Определение мультиколлинеарности

Мультиколлинеарность — наличие линейной зависимости между регрессорами

- строгая мультиколлинеарность — идеальная линейная зависимость
- нестрогая мультиколлинеарность — примерная линейная зависимость

Строгая мультиколлинеарность

Строгая мультиколлинеарность — идеальная линейная зависимость между регрессорами.
Пример:

$$X = \begin{pmatrix} 1 & 4 & 12 & 8 \\ 1 & 3 & 3 & 3 \\ 1 & 1 & 7 & 4 \\ 1 & 2 & 4 & 3 \\ 1 & 3 & 5 & 4 \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

Здесь: $x_{\cdot 2} + x_{\cdot 3} = 2x_{\cdot 4}$

Строгая мультиколлинеарность

Частая причина: неправильно включены дамми-переменные
Пример с ошибкой:

$$wage_i = \beta_1 + \beta_2 male_i + \beta_3 female_i + \beta_4 educ_i + \varepsilon_i$$

Здесь: $x_{\cdot 1} = x_{\cdot 2} + x_{\cdot 3}$

$$X = \begin{pmatrix} 1 & 1 & 0 & 16 \\ 1 & 1 & 0 & 11 \\ 1 & 0 & 1 & 18 \\ 1 & 1 & 0 & 10 \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

Последствия строгой мультиколлинеарности

В теории: оценки МНК неединственны.
Данные три модели эквивалентны:

$$\widehat{wage}_i = 15 + 3male_i - 2female_i + 3educ_i$$

$$\widehat{wage}_i = 28 - 10male_i - 15female_i + 3educ_i$$

$$\widehat{wage}_i = 18 + 0male_i - 5female_i + 3educ_i$$

Строгая мультиколлинеарность на практике

Если попросить компьютерный пакет оценить регрессию со строгой мультиколлинеарностью, то пакет может:

- выдать сообщение об ошибке
- автоматически удалить переменную [R]

Нестрогая мультиколлинеарность

Нестрогая мультиколлинеарность — примерная линейная зависимость между регрессорами

Причины:

- регрессоры, измеряющие примерно одно и то же: валютный курс на начало и на конец дня
- естественные соотношения между регрессорами: возраст, стаж и количество лет обучения

Последствия нестрогой мультиколлинеарности

Нестрогая мультиколлинеарность НЕ нарушает стандартный набор предпосылок

Оценки $\hat{\beta}_j$ несмещенные, асимптотически нормальные, можно проверять гипотезы и строить доверительные интервалы

Последствия нестрогой мультиколлинеарности

Хотя бы один из регрессоров хорошо объясняется другими регрессорами

$$se^2(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{RSS_j} = \frac{\hat{\sigma}^2}{TSS_j \cdot (1 - R_j^2)} = \frac{1}{1 - R_j^2} \frac{\hat{\sigma}^2}{TSS_j}$$

Высокие стандартные ошибки $se(\hat{\beta}_j)$

Неприятные проявления высоких стандартных ошибок

- очень широкие доверительные интервалы
- незначимые коэффициенты
- чувствительность модели к добавлению/удалению наблюдения

Практический признак нестрогой мультиколлинеарности

На практике мультиколлинеарность можно заметить, если:

- Несколько коэффициентов незначимы по отдельности
- Гипотеза об их одновременном равенстве нулю отвергается

Количественные признаки мультиколлинеарности

- коэффициент вздутия дисперсии (Variance Inflation Factor)

$$VIF_j = \frac{1}{1 - R_j^2}$$

$$se^2(\hat{\beta}_j) = VIF_j \cdot \frac{\hat{\sigma}^2}{TSS_j}$$

- выборочные корреляции между регрессорами

Некоторые источники: $VIF_j > 10$, $sCorr(x, z) > 0.9$

Что делать?

- Не так страшен чёрт! Оценки $\hat{\beta}_j$ обладают наименьшей дисперсией среди несмещенных оценок. На доверительных интервалах для прогнозов мультиколлинеарность не сказывается.
- Немного пожертвовать несмещенностью, чтобы сильно уменьшить дисперсию
- Мечта: получить больше наблюдений

Жертвуем несмещенностью

В модели $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \ldots + \varepsilon_i$ есть мультиколлинеарность.

- выкинуть часть регрессоров

Жертвуем: знанием выкидываемого коэффициента, несмещенностью оставшихся коэффициентов.

- использовать МНК со штрафом

Жертвуем: несмещенностью коэффициентов, доверительными интервалами.

Упражнение [у доски]

$y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \beta_4 w_i + \varepsilon_i$
$R_2^2 = 0.5,\ R_3^2 = 0.95,\ R_4^2 = 0.98$

- Рассчитайте коэффициенты вздутия дисперсии
- Имеет ли место нестрогая мультиколлинеарность
- Между какими переменными есть существенная линейная зависимость?

МНК со штрафом

Общая идея МНК со штрафом:

- Обычный МНК

$$RSS \to \min$$

- МНК со штрафом

$$RSS + \lambda \cdot (\text{суммарный размер всех коэффициентов}) \to \min$$

Три популярных варианта оштрафовать МНК

- Ридж-регрессия

$$\min_{\hat{\beta}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{k} \hat{\beta}_j^2$$

- LASSO

$$\min_{\hat{\beta}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{k} |\hat{\beta}_j|$$

- Метод эластичной сети

$$\min_{\hat{\beta}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda_1 \sum_{j=1}^{k} |\hat{\beta}_j| + \lambda_2 \sum_{j=1}^{k} \hat{\beta}_j^2$$

Упражнение [у доски]

Выведите оценку $\hat{\beta}_{Ridge}$ в модели $y_i = \beta x_i + \varepsilon_i$

Метод главных компонент

Идея:
Позволяет уменьшить число переменных, выбрав самые изменчивые.
Подробности:

- Из старых переменных создаются новые переменные.
- Новые переменные (главные компоненты) являются линейными комбинациями старых.
- Исходные переменные предварительно центрируются (то есть из каждой переменной вычитается её среднее значение).

Переход к новым переменным

Например:
Исходные переменные (центрированные): x_1 и x_2
Новые переменные (главные компоненты):
$pc_1 = \frac{1}{\sqrt{2}}x_1 + \frac{1}{\sqrt{2}}x_2$
$pc_2 = \frac{1}{2}x_1 - \frac{\sqrt{3}}{2}x_2.$
Сумма квадратов весов, с которыми исходные переменные входят в каждую новую, равна 1.

Новые переменные

- pc_1 имеет максимальную выборочную дисперсию $sVar(pc_1)$
- pc_2 некоррелирована с pc_1 и имеет максимальную $sVar(pc_2)$
- pc_3 некоррелирована с pc_1, pc_2 и имеет максимальную $sVar(pc_3)$
- ...

Игрушечный пример для пояснения идеи

Биология	Математика
4	5
4	2
4	5
4	4
4	3
4	4
3	3
5	3

Упрощенно:

Первая главная компонента — математика

Вторая главная компонента — биология

Упражнение [у доски]

Найдите первую главную компоненту

a_1	a_2
2	5
4	1
0	3

Не забываем центрировать!

Свойства главных компонент

$pc_1 = v_{11} \cdot x_1 + v_{21} \cdot x_2 + \ldots + v_{k1} \cdot x_k$

$\ldots$

$pc_k = v_{1k} \cdot x_1 + v_{2k} \cdot x_2 + \ldots + v_{kk} \cdot x_k$

$$sCorr(pc_j, pc_m) = 0$$

$$sVar(x_1)+sVar(x_2)+\ldots+sVar(x_k) = sVar(pc_1)+sVar(pc_2)+\ldots+sVar(pc_k)$$

Немного про линейную алгебру главных компонент

Если: все переменные центрированы, $\bar{x}_j = 0$
То: $pc_j = X \cdot v_j$ и $|pc_j|^2 = \lambda_j$, где
λ_j — собственные числа, а v_j — собственные вектора матрицы $X'X$

Что дают главные компоненты?

- визуализировать сложный набор данных
- увидеть самые информативные переменные
- увидеть особенные наблюдения
- переход к некоррелированным переменным

Подводные камни на практике

- разные единицы измерения
- применение перед регрессией

Разные единицы измерения

Первая главная компонента “поймает” переменную с самыми мелкими единицами измерения.

Вместо самой информативной переменной первой компонентой станет самая шумная.

Решение:

Нормировать переменные перед применением метода главных компонент:

$x_j = \frac{a_j - \bar{a}_j}{sd(a_j)}$

Применение перед регрессией

Очень часто строят регрессию на несколько первых главных компонент, например на pc_1, pc_2.
Осторожно:
Хорошо объясняющая переменная может быть почти постоянной.
Подход применим, но надо быть уверенным, что сильная изменчивость регрессора статистически связана с зависимой переменной.

Метод главных компонент

- прежде всего полезен сам по себе (!)
- иногда используется для борьбы с мультиколлинеарностью

Мораль - мультиколлинеарность

- Мультиколлинеарность — нестрогая линейная зависимость между регрессорами
- Основное последствие: высокие стандартные ошибки. Следовательно, широкие доверительные интервалы для коэффициентов.
- Либо не бороться, либо жертвовать несмещенностью.
- LASSO, Ridge — два варианта МНК со штрафом