Мультиколлинеарность

Эконометрика. Лекция 4

Лекция 4. Мультиколлинеарность.

Мультиколлинеарность — наличие линейной зависимости между регрессорами.

- строгая (идеальная линейная зависимость)
- нестрогая (примерная линейная зависимость)

Строгая мультиколлинеарность

Пример:

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

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

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

Последствия строгой мультиколлинеарности

в теории: оценки МНК неединственны

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

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

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

на практике:

- сообщение об ошибке
- автоматическое удаление переменной, R

Нестрогая мультиколлинеарность

Причина:

- 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

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 проявление высоких стандартных ошибок

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

Типичное проявление

Несколько коэффициентов незначимы по отдельности

Гипотеза об их одновременном равенстве нулю отвергается.

количественные признаки

- 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: $VIF_j > 10$, $\widehat{Corr}(x_{\cdot j}, x_{\cdot m}) > 0.9$

Что делать?

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

Жертвуем несмещенностью

Модель зависит от всех регрессоров!

- выкинуть часть регрессоров

Жертвуем: знанием коэффициента, несмещенностью коэффициентов

- использовать МНК со штрафом

Жертвуем: несмещенностью коэффициентов, доверительными интервалами

Жертвуем несмещенностью!

упражнение у чудо доски:

$R_2^2 = 0.5$, $R_3^2 = 0.95$, $R_4^2 = 0.98$

Рассчитайте VIF_j, между какими переменными есть линейная зависимость?

МНК со штрафом

- 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

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

чудо-доска, упражнение

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

метод главных компонент

Позволяет уменьшить число переменных, выбрав самые изменчивые

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 равна 1.

Новые переменные

- pc_1 имеет максимальную выборочную дисперсию $\widehat{Var}(pc_1)$
- pc_2 некоррелирована с pc_1 и имеет максимальную $\widehat{Var}(pc_2)$
- pc_3 некоррелирована с pc_1, pc_2 и имеет максимальную $\widehat{Var}(pc_3)$
- ...

игрушечный пример для пояснения идеи

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

Первая главная компонента — математика

Вторая главная компонента — биология

чудо-доска

Найдите первую главную компоненту

a_1	a_2
2	5
4	1
0	3

Не забываем центировать!

Свойства главных компонент

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

$$\ldots$$

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

$$\widehat{Corr}(pc_j, pc_m) = 0$$

$$\widehat{Var}(x_1) + \widehat{Var}(x_2) + \ldots + \widehat{Var}(x_k) = \widehat{Var}(pc_1) + \widehat{Var}(pc_2) + \ldots + \widehat{Var}(pc_k)$$

Вставка с линейной алгеброй

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 <> переменную с самыми мелкими единицами измерения

вместо самой информативной — самая шумная

нормировать переменные $x_j = \frac{a_j - \bar{a}_j}{se(a_j)}$

Применение перед регрессией

строят регрессию на несколько первых главных компонент, например на pc_1, pc_2

Осторожно:

хорошо объясняющая переменная может быть почти постоянной

Метод главных компонент

- полезен сам по себе
- иногда используется для борьбы с мультиколлинеарностью

Мораль - мультиколлинеарность

- зависимость между регрессорами
- высокие стандартные ошибки
- либо не бороться, либо жертвовать несмещенностью