Лекция 4. Мультиколлинеарность.

1. Мультиколлинеарность — наличие линейной зависимости между регрессорами.

   (a) строгая (идеальная линейная зависимость)

   (b) нестрогая (примерная линейная зависимость)

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

3. 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: $x_{.1} = x_{.2} + x_{.3}$

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

4. Последствия строгой мультиколлинеарности

в теории: оценки МНК неединственны

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

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

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

5. на практике:
   (a) сообщение об ошибке
   (b) автоматическое удаление переменной (`R`)
6. Нестрогая мультиколлинеарность

   Причина:

   (a) регрессоры, измеряющие примерно одно и то же: валютный курс на начало и на конец дня
   (b) естественные соотношения между регрессорами: возраст, стаж и количество лет обучения
7. последствия нестрогой мультиколлинеарности

нестрогая мультиколлинеарность НЕ нарушает стандартный набор предпосылок

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

8. последствия

   один из регрессоров хорошо объясняется другими регрессорами

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

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

9. неприятные проявление высоких стандартных ошибок

   (a) очень широкие доверительные интервалы

   (b) незначимые коэффициенты

   (c) чувствительность модели к добавлению/удалению наблюдения

10. Типичное проявление

    Несколько коэффициентов незначимы по отдельности

    Гипотеза об их одновременном равенстве нулю отвергается.

11. количественные признаки

    (a) коэффициент вздутия дисперсии (Variance Inflation Factor)

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

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

    (b) выборочные корреляции между регрессорами

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

12. Что делать?

(a) Не так страшен чёрт! Оценки $\hat{\beta}_j$ обладают наименьшей дисперсией среди несмещенных оценок. На доверительных интервалах для прогнозов мультиколлинеарность не сказывается.

(b) Пожертвовать несмещенностью

(c) Мечта: получить больше наблюдений

13. Жертвуем несмещенностью

Модель зависит от всех регрессоров!

(a) выкинуть часть регрессоров

Жертвуем: знанием коэффициента, несмещенностью коэффициентов

(b) использовать МНК со штрафом

Жертвуем: несмещенностью коэффициентов, доверительными интервалами

Жертвуем несмещенностью!

14. 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?

15. МНК со штрафом

    (a) Ридж-регрессия

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

    (b) LASSO

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

(c) Метод эластичной сети

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

16. чудо-доска, упражнение

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

17. метод главных компонент

    Позволяет уменьшить число переменных, выбрав самые изменчивые

18. 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.

19. Новые переменные

    (a) pc_1 имеет максимальную выборочную дисперсию $\widehat{Var}(pc_1)$

    (b) pc_2 некоррелирована с pc_1 и имеет максимальную $\widehat{Var}(pc_2)$

    (c) pc_3 некоррелирована с pc_1, pc_2 и имеет максимальную $\widehat{Var}(pc_3)$

    (d) ...

20. 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

21. чудо-доска

Найдите первую главную компоненту

a_1	a_2
3	0
3	4
0	5

Не забываем центировать!

22. Свойства главных компонент

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

$\ldots$

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

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

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

23. 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 $X'X$

24. Что дают главные компоненты?

    (a) визуализировать сложный набор данных
    (b) увидеть самые информативные переменные
    (c) увидеть особенные наблюдения
    (d) переход к некоррелированным переменным

25. Подводные камни на практике

    (a) разные единицы измерения
    (b) применение перед регрессией

26. Разные единицы измерения

    первая главная компонента «поймает» переменную с самыми мелкими единицами измерения вместо самой информативной — самая шумная

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

27. Применение перед регрессией

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

    Осторожно:

    хорошо объясняющая переменная может быть почти постоянной

28. Метод главных компонент

    (a) полезен сам по себе

    (b) иногда используется для борьбы с мультиколлинеарностью

29. Мораль - мультиколлинеарность

    (a) зависимость между регрессорами

    (b) высокие стандартные ошибки

(c) либо не бороться, либо жертвовать несмещенностью