NATIONAL RESEARCH UNIVERSITY

Лекция 03.12.21, 2 модуль

Мультиколлинеарность

Демидова О.А.

E-mail: demidova@hse.ru

План лекции

- Идеальная и практическая мультиколлинеарность (квазимультиколлинеарность).
- Последствия мультиколлинеарности
- Признаки наличия мультиколлинеарности
- Показатели степени мультиколлинеарности
- Методы борьбы с мультиколлинеарностью

Мультиколлинеарность

Теоретическая мультиколлинеарность данных – явление, наблюдаемое при нарушении условий теоремы Гаусса – Маркова об отсутствии точной линейной связи между регрессорами, т.е. для модели $Y = X\beta + \varepsilon$

$rangX <$ число столбцов в матрице X.

При наличии теоретической мультиколлинеарности однозначное нахождение оценов МНК коэффициентов регрессии невозможно, система нормальных уравнений $X'X\beta = X'Y$ не имеет единственного решения.

Теоретическая мультиколлинеарность

$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_K X_K + \varepsilon,$

Теоретическая мультиколлинеарность:

$\mathrm{Rank}(X) < k + 1$

Пример 1. $\ln w\, age = \beta_0 + \beta_1 S + \beta_2 MALE + \beta_3 FEMALE +$

$\cdots + \varepsilon,$

$FEMALE + MALE = I$, I – единичный вектор

Пример 2. $ln\, price = \beta_0 + \beta_1 livsq + \beta_2 nonlivsq + \beta_3 totsq +$

$\cdots + \varepsilon,$

$livsq + nonlivsq = totsq$

Квазимультиколлинеарность

При работе с реальными данными часто имеет место квазимультиколлинеарность, когда между регрессорами существует почти линейная зависимость.

$Y = X\beta + \varepsilon$

теоретическая мультиколлинеарность:

$\text{rang}X < k + 1$

$\Rightarrow \text{rang}(X'X) = \text{rang}X < k + 1 \Rightarrow \det(X'X) = 0.$

Квазимультиколлинеарность:

$\det(X'X) \approx 0$

Но это зависит от единиц, в которых измеряются переменные!

Последствия мультиколлинеарности

$$\hat{\beta} = (X'X)^{-1}X'Y,$$

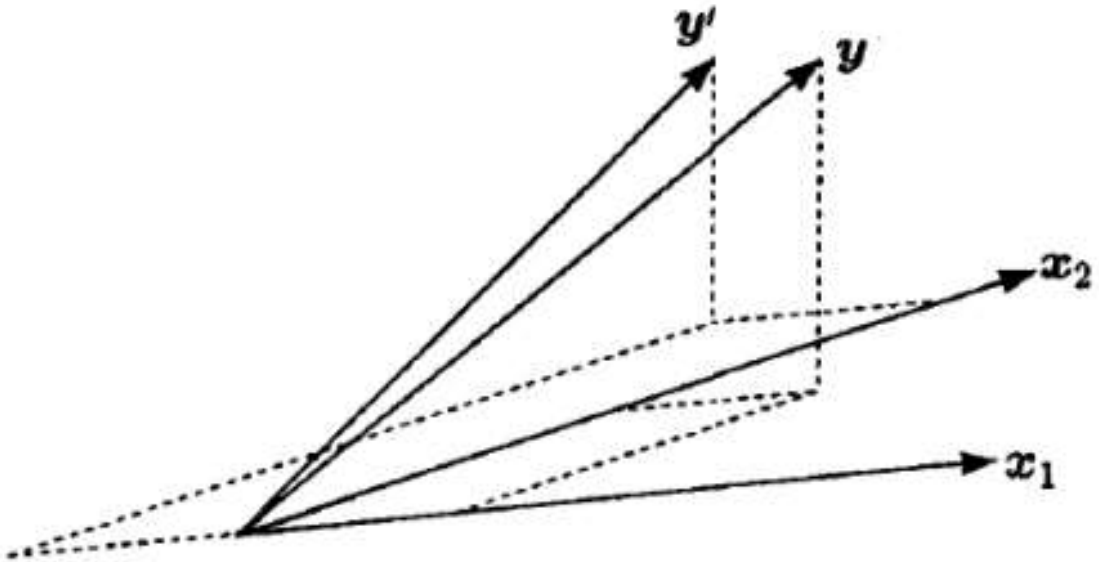
$$Var(\hat{\beta}) = \sigma_{\varepsilon}^{2}(X'X)^{-1},$$

$$A^{-1} = \frac{1}{\det A}A^{*}$$

Нестабильность оценок параметров регрессии и их дисперсий при малых изменениях исходных данных в случае мультиколлинеарности

Геометрическая интерпретация

Векторы y и y' мало отличаются друг от друга, но в силу того, что угол между векторами (регрессорами) x_1 и x_2 мал, разложения проекций этих двух веторов по x_1 и x_2 отличается значительно.



Источник: Магнус, Катышев, Пересецкий (2007).

Мультиколлинеарность, пример

$$y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon,$$

$$x_1'x_1 = x_2'x_2 = 1,\ \sigma_\varepsilon^2 = 1,$$

$$x_1'x_2 = x_2'x_1 = r,$$

$$\mathrm{var}(\hat{\beta}) = \sigma_\varepsilon^2 (x'x)^{-1},$$

$$x'x = \begin{pmatrix} x_1'x_1 & x_1'x_2 \\ x_2'x_1 & x_2'x_2 \end{pmatrix} = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix},$$

$$\mathrm{var}(\hat{\beta}_1) = \mathrm{var}(\hat{\beta}_2) = \frac{1}{1 - r^2}.$$

Признаки мультиколлинеарности

- Небольшие изменения в данных приводят к значительным изменениям в оценках коэффициентов регрессии.
- Многие коэффициенты по-отдельности не значимы, хотя в целом регрессия адекватная, R^2 может быть достаточно высоким.
- Оценки коэффициентов регрессии (обычно незначимых) могут иметь “неправильный” знак (с экономической точки зрения).

Индикаторы мультиколлинеарности

•В корреляционной матрице факторов встречаются элементы, по модулю близкие к 1.

• Достаточно большое значение
VIF – variance inflation factor хотя бы для одного фактора

$$VIF(X_j) = \frac{1}{1 - R_j^2},$$

где R_j^2 – коэффициент множественной детерминации регрессора Xj на все остальные регрессоры,

$$(Var(\hat{\beta}_j) = \frac{\sigma_\varepsilon^2}{TSS_j(1 - R_j^2)},\ TSS_j = (X_j - \bar{X}_j \mathrm{I})'(X_j - \bar{X}_j \mathrm{I})).$$

Индикаторы мультиколлинеарности

CN (conditional number) – число обусловленности

Матрица X (n x (k+1)) не является квадратной, $rangX = k + 1$.

Матрица $X'X$ ((k+1) x (k+1)) – положительно определенная квадратная матрица, у нее существует k+1 собственных значений (все положительные): $\lambda_1, \dots, \lambda_{k+1}$.

$CN(X) = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$, если этот показатель > 30, то это может свидетельствовать о мультиколлинеарности.

В некоторых статистических пакетах используют в качестве определения $CN = \frac{\lambda_{max}}{\lambda_{min}}$.

Пример мультиколлинеарности данных

$$EARNINGS = \beta_0 + \beta_1 S + \beta_2 SVABC_1 + \beta_3 ASVABC_2 + \beta_4 ASVABC_3 + \beta_5 ASVABC_4 + \varepsilon$$

```
reg  EARNINGS        S ASVAB01 ASVAB02 ASVAB03 ASVAB04
Source      SS              df      MS          Number of obs        =     540
                                                F(  5,   534) =   30.60
Model       24945.2724       5  4989.05448      Prob > F     =  0.0000
Residual    87064.9587     534  163.042994      R-squared    =  0.2227
                                                Adj R-squared =  0.2154
Total       112010.231     539  207.811189      Root MSE     =  12.769

EARNINGS       Coef.        Std. Err.      t     P>t        [95% Conf.Interval]

S              1.700556     .2781761     6.11   0.000       1.154102   2.247009
ASVAB01         .0640055    .0997875     0.64   0.522      -.1320188   .2600297
ASVAB02         .4385383    .091164      4.81   0.000       .2594542   .6176223
ASVAB03        -.1433842    .1202383    -1.19   0.234      -.3795824   .0928139
ASVAB04        -.0265344    .0985583    -0.27   0.788      -.2201438   .1670751
_cons        -20.48614      3.600184    -5.69   0.000      -27.5584   -13.41388
```

Много незначимых коэффициентов

Пример мультиколлинеарности данных

vif

Variable	VIF	1/VIF
ASVAB03	4.20	0.238017
ASVAB04	3.01	0.332532
ASVAB01	3.00	0.333805
ASVAB02	2.64	0.378371
S	1.52	0.657411
Mean VIF	2.87	

Методы борьбы с мультиколлинеарностью

- **Переспецификация модели (функциональные преобразования переменных)**
- **Исключение одной или нескольких объясняющих переменных**
- **Метод главных компонент**
- **Использование ridge (гребневых), LASSO и т.п. оценок параметров**

Метод главных компонент

Хотим перейти от факторов $X_1, \dots, X_k$ к факторам, которые не коррелируют.

$X = [X_1, \dots, X_k],\ Var[X] = V,$

$Z_1 = \alpha_{11}X_1 + \cdots \alpha_{1k}X_k = \alpha_1{}'X, \quad \alpha_{11}^2 + \cdots + \alpha_{1k}^2 = 1,$

$V\,\mathrm{ar}(Z_1) \to \max,\ \alpha_1{}'\alpha_1 = 1,$

$Var(Z_1) = \alpha_1{}'V\alpha_1,$

$L(\alpha_1) = \alpha_1{}'V\alpha_1 - \lambda_1(\alpha_1{}'\alpha_1 - 1) \to \max,$

$V\alpha_1 = \lambda_1\alpha_1,$

λ_1 – максимальное собственное значение матрицы V,

α_1 – соответствующий собственный вектор.

Метод главных компонент

Пусть $\lambda_1, \dots, \lambda_k$ – собственные значения матрицы V (убывающие),

$\alpha_1, \dots, \alpha_k$ – соответствующие собственные векторы.

$Z_1 = \alpha_1{}'X, \dots, Z_k = \alpha_k{}'X.$

$Var(Z_1) = \alpha_1{}'V\alpha_1 = \lambda_1, \dots, Var(Z_k) = \alpha_k{}'V\alpha_k = \lambda_k.$

$Z_1, \dots, Z_k$ называются главными компонентами для $X_1, \dots, X_k$.

Свойства главных компонент:

1) $\alpha_1, \dots, \alpha_k$ являются ортогональными по свойству собственных векторов $\Rightarrow Z_1, \dots, Z_k$ ортогональны (= не коррелируют), т.к.

$Z'_m Z_l = (\alpha'_m \mathrm{X})' \, (\alpha'_l \mathrm{X}) = X' \, (\alpha_m \, \alpha'_l)\mathrm{X} = 0$

Метод главных компонент

Свойства главных компонент:

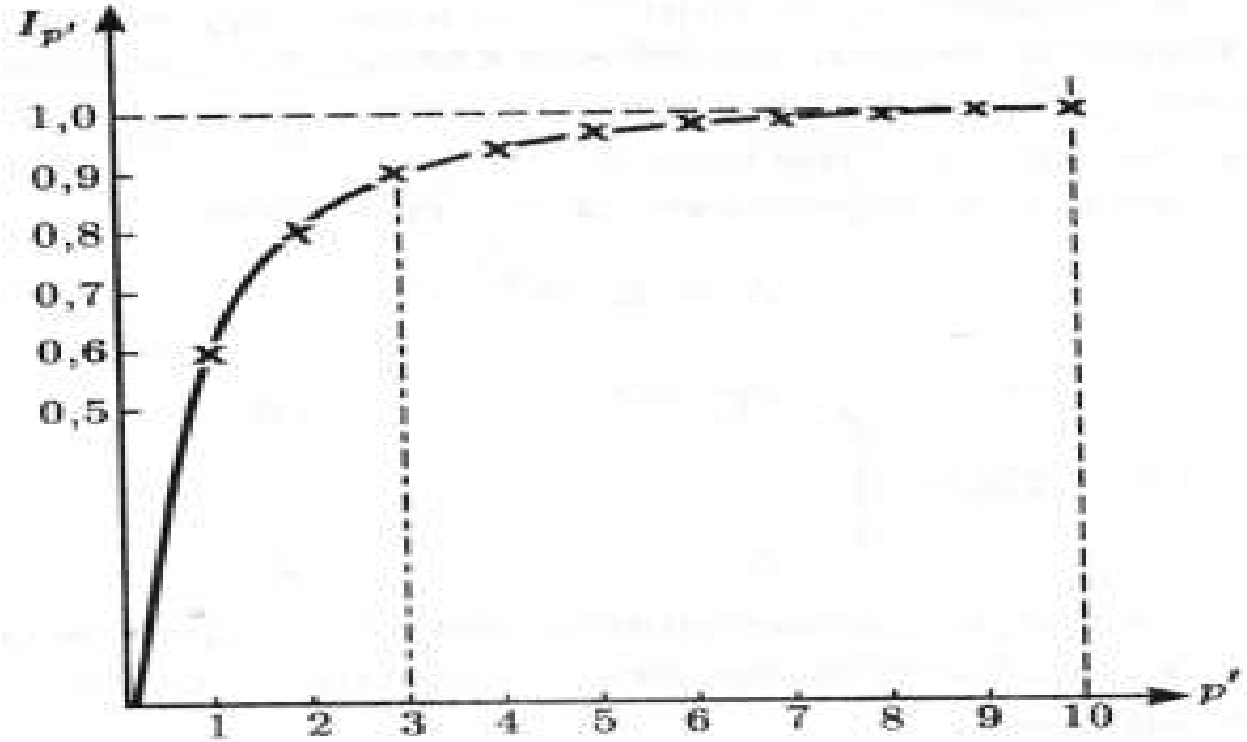
2) $Var(Z_1) + \cdots + Var(Z_k) = \lambda_1 + \cdots + \lambda_k =$

$= trace(V) = Var(X_1) + \cdots + Var(X_k)$

Доля общей дисперсии $X_1,\ldots, X_k$, которую объясняют m главных компонент, равна $\frac{\lambda_1+\cdots+\lambda_m}{\lambda_1+\cdots+\lambda_k}$

Пример

Доля общей дисперсии признаков, объясняемая одной, двумя, тремя и тд главными компонентами



Изменение относительной доли суммарной дисперсии исследуемых признаков, обусловленной первыми p' главными компонентами, в зависимости от p' (случай $p = 10$)

Примеры

Примечание: Если все переменные измеряются в одних единицах, то для нахождения главных компонент используют ковариационную матрицу.

Если же факторы измеряют в разных единицах, то их линейная комбинация не имеет смысла, поэтому факторы сначала нормируют (делят на корень из выборочной дисперсии), тогда ковариационная матрица превращается в корреляционную матрицу.

Примеры

1) См. примеры 4.1, 4.2 из книги «Прикладная статистика в задачах и упражнениях», авторы С.А.Айвазян, В.С.Мхитарян

2) См. кластеризацию регионов в пространстве двух главных компонент в статье «Метод кластеризации регионов РФ с учетом отраслевой структуры ВРП», авторы С. А. Айвазян, М. Ю. Афанасьев, А. В. Кудров

Пример 4.1

4.Б. Примеры решения типовых задач и упражнений

Пример 4.1 (задача). При формировании типообразующих признаков предприятий отрасли были обследованы 24 предприятия ($n = 24$) по трем технико-экономическим показателям: объему выпускаемой продукции $x^{(1)}$ (тыс.условных денежных единиц), основным фон-

дам $x^{(2)}$ (тыс.у.д.е.) и фонду оплаты труда $x^{(3)}$ (тыс.у.д.е.). По полученным в результате обследования исходным статистическим данным $(x_i^{(1)}, x_i^{(2)}, x_i^{(3)})$, — $i = 1, 2, \ldots, 24$, — были получены оценки вектора средних значений $\hat{\mathbf{a}} = (\hat{a}^{(1)}, \hat{a}^{(2)}, \hat{a}^{(3)})^\top = (420;\ 240;\ 85)^\top$ и ковариационной матрицы

$$\widehat{\Sigma} = \begin{pmatrix} 451{,}39 & 271{,}17 & 168{,}70 \\ 271{,}17 & 171{,}73 & 103{,}29 \\ 168{,}70 & 103{,}29 & 66{,}65 \end{pmatrix}$$

Требуется:

1) *вывести уравнения для вычисления главных компонент* $z^{(1)}, z^{(2)}$ *и* $z^{(3)}$ *по заданным значениям исходных технико-экономических показателей* $x^{(1)}, x^{(2)}$ *и* $x^{(3)}$;
2) *определить относительные доли суммарной дисперсии, обусловленные одной и двумя главными компонентами;*

Пример 4.1

4.Б. Примеры решения типовых задач и упражнений

Пример 4.1 (задача). При формировании типообразующих признаков предприятий отрасли были обследованы 24 предприятия ($n = 24$) по трем технико-экономическим показателям: объему выпускаемой продукции $x^{(1)}$ (тыс.условных денежных единиц), основным фон-

дам $x^{(2)}$ (тыс.у.д.е.) и фонду оплаты труда $x^{(3)}$ (тыс.у.д.е.). По полученным в результате обследования исходным статистическим данным $(x_i^{(1)}, x_i^{(2)}, x_i^{(3)})$, — $i = 1, 2, \ldots, 24$, — были получены оценки вектора средних значений $\hat{\mathbf{a}} = (\hat{a}^{(1)}, \hat{a}^{(2)}, \hat{a}^{(3)})^\top = (420;\ 240;\ 85)^\top$ и ковариационной матрицы

$$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 451{,}39 & 271{,}17 & 168{,}70 \\ 271{,}17 & 171{,}73 & 103{,}29 \\ 168{,}70 & 103{,}29 & 66{,}65 \end{pmatrix}$$

Требуется:

1) *вывести уравнения для вычисления главных компонент* $z^{(1)}, z^{(2)}$ *и* $z^{(3)}$ *по заданным значениям исходных технико-экономических показателей* $x^{(1)}, x^{(2)}$ *и* $x^{(3)}$;
2) *определить относительные доли суммарной дисперсии, обусловленные одной и двумя главными компонентами;*

Пример 4.1

Решение

1) Для определения коэффициентов линейного преобразования (4.5), с помощью которого осуществляется переход к главным компонентам, необходимо решить вначале характеристическое уравнение (4.8), а затем использовать найденные собственные значения $\lambda_1, \lambda_2, \ldots, \lambda_p$ для подстановки в системы уравнений (4.7), решения которых и дают коэффициенты $l_j = (l_{j1}, l_{j2}, \ldots, l_{jp})$. В данной задаче

$$|\widehat{\Sigma} - \lambda I| = \begin{vmatrix} 451{,}39 - \lambda & 217{,}17 & 168{,}70 \\ 271{,}17 & 171{,}73 - \lambda & 103{,}29 \\ 168{,}70 & 101{,}29 & 66{,}65 - \lambda \end{vmatrix} = 0,$$

откуда находим $\lambda_1 = 680{,}40$, $\lambda_2 = 6{,}50$, $\lambda_3 = 2{,}86$.

Последовательно подставляя эти значения в систему (4.7) и решая эти системы относительно $l_j = (l_{j1}, l_{j2}, l_{j3})$, $-j = 1, 2, 3$, — получаем: $l_1 = (0{,}813; 0{,}496; 0{,}307)$; $l_2 = (-0{,}545; 0{,}832; 0{,}101)$, $l_3 = (-0{,}205; -0{,}249; 0{,}947)$, так что уравнения для вычисления главных компонент $z^{(j)}$ $(j = 1, 2, 3)$ будут иметь вид:

$$\begin{aligned} z^{(1)} &= 0{,}81(x^{(1)} - 420) + 0{,}50(x^{(2)} - 240) + 0{,}31(x^{(3)} - 85), \\ z^{(2)} &= -0{,}55(x^{(1)} - 420) + 0{,}83(x^{(2)} - 240) + 0{,}10(x^{(3)} - 85), \\ z^{(3)} &= -0{,}21(x^{(1)} - 420) - 0{,}25(x^{(2)} - 240) + 0{,}95(x^{(3)} - 85). \end{aligned} \qquad (4.20)$$

Пример 4.1

2) В соответствии с равенствами (4.10) и (4.10′) и вытекающим из них представлением $I_{p'}(Z)$ в виде (4.4′) имеем:

$$I_1(Z(X)) = \frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3} = 0{,}9864;$$

$$I_2(Z(X)) = \frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3} = 0{,}9958;$$

что свидетельствует о том, что почти вся информация (а именно, 98,64%) о специфике предприятия данного типа, описанной с помощью переменных $x^{(1)}, x^{(2)}$ и $x^{(3)}$, *содержится в одной лишь первой главной компоненте* $x^{(1)}$.

Пример 4.2

Пример 4.2 (упражнение). Компонентный анализ проведен по данным двадцати сельскохозяйственных районов ($n = 20$) области, которые содержат результаты измерений следующих показателей: $x^{(1)}$ — число колесных тракторов на 100 га; $x^{(2)}$ — число зерноуборочных комбайнов на 100 га; $x^{(3)}$ — число орудий поверхностной обработки почвы на 100 га; $x^{(4)}$ — количество удобрений, расходуемых на гектар; $x^{(5)}$ — количество средств защиты растений, расходуемых на гектар (исходные статистические данные $x_i^{(1)}, x_i^{(2)}, x_i^{(3)} x_i^{(4)}, x_i^{(5)}$, — $i = 1, 2, \ldots, 20$, — приведены в табл. П2.1 Приложения 2).

Расчеты проводились по *нормированным данным* вида (4.1′) и представлены в следующей таблице:

Главные компоненты $x^{(i)}$	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$
Собственные значения λ_i	3,04	1,41	0,43	0,10	0,02
Вклад i-й главной компоненты (%) в суммарную дисперсию	60,8	28,2	8,6	2,0	0,4
Суммарный вклад первых главных компонент (%)	60,8	89,0	97,6	99,6	100,0

При расчете относительного вклада главных компонент учитывалось, что $\sum_{i=1}^{p} \lambda_i = p = 5$. Для анализа были оставлены две первые главные компоненты ($p' = 2$), на которые приходится 89% суммарной вариации.

Для интерпретации главных компонент построена матрица факторных нагрузок

$$\mathbf{A} = \begin{pmatrix} 0{,}95^{*} & 0{,}97^{*} & 0{,}94^{*} & 0{,}24 & 0{,}56 \\ -0{,}19 & -0{,}17 & -0{,}28 & 0{,}88^{*} & 0{,}67^{*} \end{pmatrix}^{\top}.$$

Пример 4.2

Звездочкой (*) отмечены элементы, удовлетворяющие условию $|a_{ij}| > 0{,}6$, т. е. те, которые следует учитывать при интерпретации главных компонент $z^{(1)}$ и $z^{(2)}$.

Требуется:

дать содержательную интерпретацию первых двух главных компонент.

Решение

Из вида матрицы A следует, что первая главная компонента наиболее тесно связана с показателями: $x^{(1)}$ — число колесных тракторов ($a_{11} = r(x^{(1)}, z^{(1)}) = 0{,}95$); $x^{(2)}$ — число зерноуборочных комбайнов ($a_{21} = r(x^{(2)}, z^{(1)}) = 0{,}97$); $x^{(3)}$ — число орудий поверхностей обработки почвы на 100 га ($a_{31} = r(x^{(3)}, z^{(1)}) = 0{,}94$). Поэтому *первая главная компонента* $z^{(1)}$ *интерпретирована как уровень механизации работ.*

Вторая главная компонента $z^{(2)}$ тесно связана с количествами удобрения ($x^{(4)}$) и средств защиты растений ($x^{(5)}$), расходуемых на гектар ($a_{42} = r(x^{(4)}, z^{(2)}) = 0{,}88$; $a_{52} = r(x^{(5)}, z^{(2)}) = 0{,}67$). Соответственно $z^{(2)}$ *интерпретируется как уровень химизации растениеводства.*

Спасибо за внимание!

Демидова О.А.
E-mail: demidova@hse.ru