Фиктивные переменные в регрессионных моделях

Лекция

Цели лекции

- Линейные регрессионные моделями с переменной структурой
- Фиктивные переменные сдвига и наклона
- Тест Чоу на наличие структурного сдвига

Необходимость использования фиктивных переменных

На практике часто возникает необходимость использования качественных признаков. Влияние качественного фактора выражают в виде фиктивной (искусственной) переменной, отражающей его два противоположных состояния:

$$D = \begin{cases} 0, & \textit{фактор не действует} \\ 1, & \textit{фактор действует} \end{cases}$$

Фиктивные переменные позволяют отразить в модели эффекты сдвига и наклона в результате воздействия качественных факторов на зависимую переменную

Примеры фиктивных переменных

Исследуется зависимость между доходом и потреблением с учетом фактора проживание (город или сельская местность)

Исследуется зависимость между продолжительностью полученного образования и доходом, и в выборке представлены как мужчины, так и женщины. Нужно выяснить, влияет ли пол на различие в результатах

Исследуется зависимость между объемом продаж магазина и средней зарплатой с учетом фактора сезонности

Пример использования фиктивной переменной

По выборочным данным (n=73) исследуется зависимость цены квартир Y на вторичном рынке жилья Санкт-Петербурга в 2000г. (тыс. долл.) от общая площадь X (м^2). Допустим мы хотим отразить в модели район – центральный или периферийный. Для этого включим в модель **фиктивную переменную сдвига *Z***: Z=0 для периферийных районов, Z=1 для центральных районов

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$$

Пример использования фиктивной переменной. Исходные данные

Исходные данные по ценам на квартиры								
X	Z	Y	X	Z	Y	X	Z	Y
30,0	1	12,0	53,7	0	15,3	81,0	0	33,0
30,3	1	12,5	54,6	0	21,0	85,0	0	32,7
31,0	0	9,5	55,0	0	21,5	87,0	1	39,0
31,0	0	11,0	55,5	1	26,0	88,0	0	34,0
32,0	1	15,9	57,0	0	21,0	90,0	1	24,5
32,2	0	12,0	57,0	1	17,0	92,0	0	23,5
33,0	0	10,5	58,0	0	17,8	92,0	0	30,0
35,0	1	14,2	60,0	1	16,5	92,5	1	43,0
35,0	1	15,6	60,0	1	17,0	93,0	0	27,0
37,0	1	16,0	62,0	0	22,0	96,0	0	38,0
38,0	0	11,0	66,0	0	23,0	96,4	0	34,0
40,0	1	23,0	66,0	0	26,0	98,0	1	32,5
42,5	0	13,5	68,0	0	16,0	100,0	0	38,6
43,0	0	14,5	68,1	1	19,5	100,0	1	30,0
44,1	1	15,2	69,7	0	23,0	105,4	1	27,3
45,0	0	14,2	70,0	0	29,0	106,0	0	41,5
45,0	0	13,3	71,0	1	24,0	107,0	1	38,0
48,0	1	22,5	74,0	0	22,0	109,0	1	42,7
48,0	1	18,5	74,0	0	23,0	110,0	0	45,5
50,0	0	18,0	74,0	0	23,0	114,3	1	35,6
50,1	0	14,0	74,7	0	26,5	114,8	1	31,0
50,6	0	16,1	75,0	0	28,0	115,0	0	37,0
53,0	0	21,0	76,4	0	28,0	116,0	1	35,0
53,0	1	15,0	79,0	0	19,5			
53,5	0	19,5	80,0	1	24,8			

Пример использования фиктивной переменной

Уравнение регрессии без учета района: $\hat{Y}_X = 1{,}797 + 0{,}3197X$

Уравнение регрессии с фиктивной переменой сдвига Z:

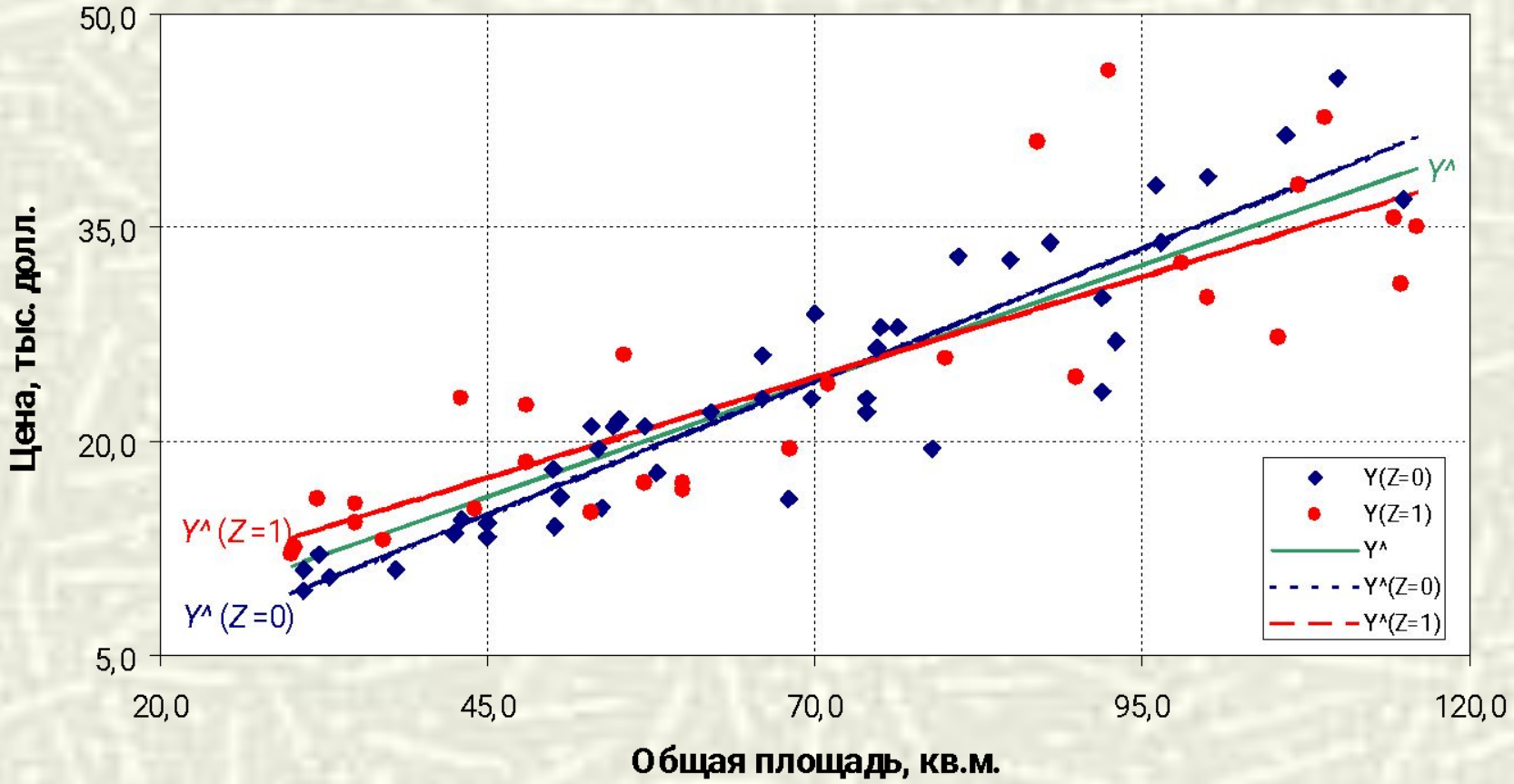
$$\hat{Y}_{X,Z} = 1{,}669 + 0{,}3193X + 0{,}3781Z = \begin{cases} 1{,}669 + 0{,}3193X, & Z = 0 \\ 2{,}047 + 0{,}3193X, & Z = 1 \end{cases}$$

Уравнения регрессии для разных районов (по частям выборки):

$\hat{Y}_0 = -1{,}942 + 0{,}3730X$ для Z=0

$\hat{Y}_1 = 5{,}323 + 0{,}2723X$ для Z=1

Пример использования фиктивной переменной. Графики



Пример использования фиктивной переменной

Из полученных уравнений и графиков видно, что одной фиктивной переменной сдвига недостаточно. Введем дополнительно фиктивную переменную, учитывающую разный наклон данных:

$$\hat{Y} = b_0 + b_1 X + (a_0 + a_1 X)Z = b_0 + b_1 X + b_2 Z + b_3 ZX$$

Учет разного сдвига

Учет разного наклона

Пример использования фиктивной переменной

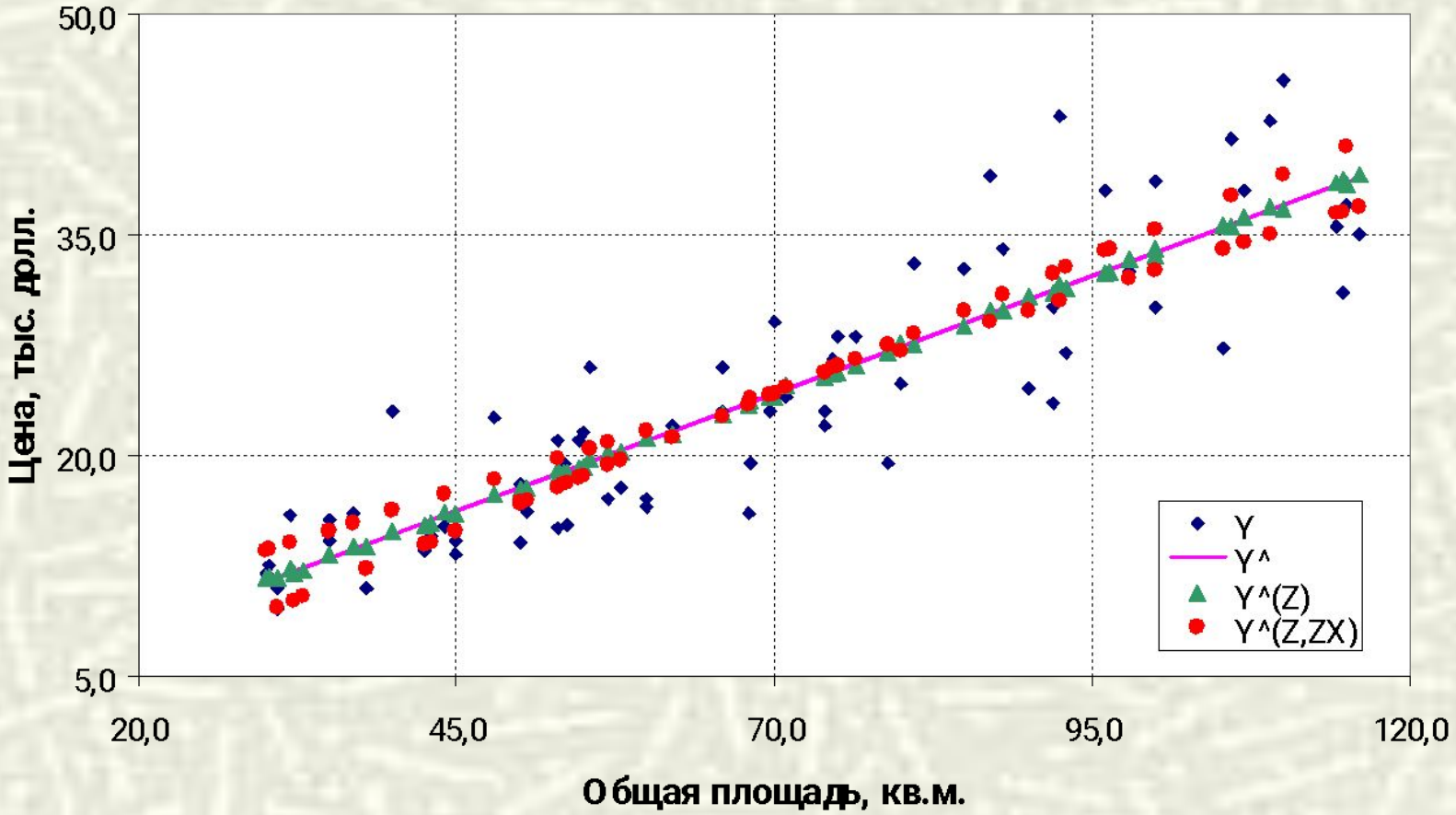
Уравнение регрессии с учетом разного сдвига *Z* и наклона *ZX*:

$$\hat{Y}_{Z,ZX} = -1{,}942 + 0{,}373X + 7{,}265Z - 0{,}1008ZX$$

или

$$\hat{Y}_{Z,ZX} = \begin{cases} -1{,}942 + 0{,}373X, & Z = 0 \\ 5{,}323 + 0{,}272X, & Z = 1 \end{cases}$$

Пример использования фиктивной переменной. Графики



Пример использования фиктивной переменной. Сравнение моделей

$$\hat{Y}_X = 1{,}797 + 0{,}3197X$$

$$\hat{Y}_0 = -1{,}942 + 0{,}373X$$

$$\hat{Y}_1 = 5{,}323 + 0{,}272X$$

$$\Leftrightarrow \quad \hat{Y}_{X,Z,ZX} = \begin{cases} -1{,}942 + 0{,}373X, & Z = 0 \\ 5{,}323 + 0{,}272X, & Z = 1 \end{cases}$$

$$\hat{Y}_{X,Z} = \begin{cases} 1{,}669 + 0{,}3193X, & Z = 0 \\ 2{,}047 + 0{,}3193X, & Z = 1 \end{cases}$$

Вывод. Уравнение регрессии с фиктивными переменными позволяет учесть в модели качественные признаки

ВЫВОД ИТОГОВ БЕЗ ФИКТИВНЫХ ПЕРЕМЕННЫХ

Регрессионная статистика	
Множественный R	0,887042419
R-квадрат	0,786844252
Нормированный R-квадрат	0,783842059
Стандартная ошибка	4,262108969
Наблюдения	73

Средняя ошибка аппроксимации: 14,6%

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1	4761,010902	4761,010902	262,0897749	1,56309E-25
Остаток	71	1289,755673	18,16557286		
Итого	72	6050,766575			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересечение	1,7967268	1,435799573	1,251377165	0,214904374
X	0,319672709	0,019746063	16,18918698	1,56309E-25

ВЫВОД ИТОГОВ С УЧЕТОМ ФИКТИВНЫХ ПЕРЕМЕННЫХ

Регрессионная статистика	
Множественный R	0,898151369
R-квадрат	0,806675881
Нормированный R-квадрат	0,798270484
Стандартная ошибка	4,117405821
Наблюдения	73

Средняя ошибка аппроксимации: 13,1%

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	3	4881,007457	1627,002486	95,97118739	1,43786E-24
Остаток	69	1169,759118	16,9530307		
Итого	72	6050,766575			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересечение	-1,941967335	1,97890111	-0,98133622	0,329855403
X	0,373030534	0,027937189	13,3524721	8,4024E-21
Z	7,264688159	2,795398543	2,598802299	0,011430154
XZ	-0,100760417	0,038272722	-2,632695374	0,010445167

Фиктивные переменные сдвига и наклона. Интерпретация коэффициентов

$$\hat{Y} = b_0 + b_1 X + b_2 Z + b_3 ZX = \begin{cases} b_0 + b_1 X, & Z = 0 \\ (b_0 + b_2) + (b_1 + b_3) X, & Z = 1 \end{cases}$$

На одной части выборки регрессия имеет коэффициенты b_0 и b_1. На другой части выборки они изменяются, соответственно, на величину коэффициентов при фиктивных переменных сдвига и наклона

Значимость коэффициентов при фиктивных переменных определяется с помощью *t*-статистики

Использование фиктивных переменных эквивалентно расчету регрессий на отдельных частях выборки

Оценка значимости влияния качественных переменных на зависимую переменную

Статистическая значимость качественных переменных проверяется по *t*-критерию: исследуем на значимость *t*-статистику коэффициента при данной фиктивной переменной

Для рассмотренного примера:

$$\left|t_{b_1}\right| = 2{,}599, \quad \left|t_{b_3}\right| = 2{,}633$$

Вывод. Район расположения квартиры значимо влияет на ее цену на уровне значимости 1% (надежность равна 99%)

Виды моделей с качественными объясняющими переменными

ANOVA-модели (модели дисперсионного анализа)
Содержат только качественные объясняющие переменные.
ANOVA-модели представляют собой кусочно-постоянные функции.

ANCOVA-модели (модели ковариационного анализа)
Содержат как количественные, так и качественные объясняющие переменные.

Использование фиктивных переменных в сезонном анализе

Учет или нейтрализация сезонного фактора с помощью фиктивных переменных

сдвига: $$Y_t = \beta_0 + \beta_1 X_t + \beta_2 D_{1t} + \beta_3 D_{2t} + \beta_4 D_{3t} + \varepsilon_t$$

сдвига и наклона: $$Y_t = \beta_0 + \beta_1 X_t + \sum_{j=1}^{3} (\beta_{j+1}^1 + \beta_{j+1}^2 X_t) D_{jt} + \varepsilon_t$$

$$D_{jt} = \begin{cases} 1, & \text{если рассматривается } (j+1)\text{-й квартал} \\ 0, & \text{в противном случае} \end{cases}$$

Фиктивная зависимая переменная. Примеры

Анализируется наличие работы у человека в зависимости от возраста, образования, семейного положения, доходов остальных членов семьи и т.д. Зависимая переменная имеет вид:

$$Y = \begin{cases} 0, & \textit{человек не имеет работы} \\ 1, & \textit{человек имеет работу} \end{cases}$$

Анализируется результат сдачи с первой попытки экзамена в ГАИ в зависимости от количества часов вождения, использования компьютерной методики обучения и т.д. Зависимая переменная:

$$Y = \begin{cases} 0, & \textit{экзамен не сдан с первой попытки} \\ 1, & \textit{экзамен сдан с первой попытки} \end{cases}$$

Фиктивная зависимая переменная. Модель и ограничения

Модель в общем случае имеет вид:

$$Y = \beta_0 + \sum_{j=1}^{m} \beta_1 X_j + \sum_{k=1}^{l} \gamma_k D_k + \varepsilon$$

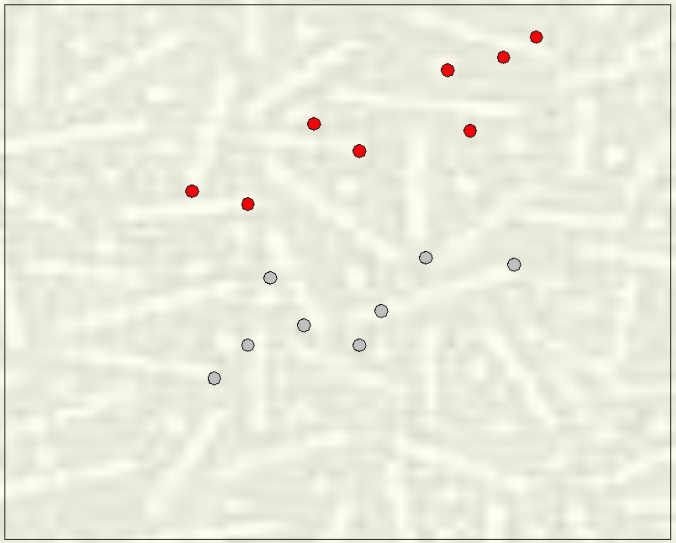
Ограниченность использования МНК для данных моделей:

1. Случайные отклонения ε_i не имеют нормального распределения.
2. Не выполняется предпосылка 2^0 постоянства дисперсии D[ε].

Для определения коэффициентов модели используют другие методы

Тест Чоу. Анализ структурных сдвигов

Пример структурного сдвига – выборка имеет две различных подвыборки



Тест Чоу. Область применения

Ситуации, когда возникает потребность в тесте Чоу:

1. Есть подозрения, что исходная выборка состоит из двух или более разных подвыборок (например, из-за различия качественной переменой)

2. К имеющийся выборке нужно присоединить дополнительные данные. И необходимо выяснить, можно ли считать обе выборки регрессионно однородными.

Суть теста Чоу: проверка гипотезы о совпадении уравнений регрессии для отдельных групп наблюдений (подвыборок)

Тест Чоу. Описание

$$F = \frac{(RSS_T - RSS_1 - RSS_2)/(m+1)}{(RSS_1 + RSS_2)/(n-2m-2)}$$

F-статистика представляет собой отношение меры улучшения качества уравнения в расчете на одну использованную степень свободы к мере необъясненной дисперсии в расчете на одну оставшуюся степень свободы

RSS_T – сумма квадратов остатков для регрессии по всей выборке; RSS_1, RSS_2 – по ее частям

Статистика имеет *F*-распределение с $(m+1)$ и $(n-2m-2)$ степенями свободы

Тест Чоу. Пример

Проверим для α=5% гипотезу о совпадении уравнений регрессии для подвыборок, соответствующих разным районам Санкт-Петербурга, из рассмотренного примера:

$RSS_T = 1289{,}756$; $RSS_1 = 509{,}179$ (Z=0); $RSS_2 = 660{,}580$ (Z+1)

$$F_{расч} = \frac{(1289{,}756 - 509{,}179 - 660{,}580)/2}{(509{,}179 + 660{,}580)/(73 - 2 - 2)} = \frac{59{,}999}{16{,}953} = 3{,}539$$

$$F_{крит} = F_{\alpha;\, m+1;\, n-2m-2} = F_{0{,}05;\, 2;\, 69} = 3{,}13 \qquad F_{расч} > F_{крит}$$

На уровне значимости 5% уравнения регрессии различны

Конец лекции