Фиктивные переменные

К теме «Множественная регрессия и корреляция»

Фиктивные переменные

На практике приходится учитывать в моделях факторы, носящие качественный характер, значения которых в наблюдениях не возможно измерить с помощью числовой шкалы.

Примеры.

Моделирование влияния пола специалистов на уровень зарплаты.

Моделирование доходов граждан от типа учебного заведения, в котором он получил образование (государственное, частное, специализированное,…)

Модель инфляции с учетом различных видов регулирования со стороны государства)

Фиктивные переменные

Возможны два подхода к решению задачи:

- построить несколько моделей отдельно для каждого значения (градации) качественной переменной
- учесть влияние качественного фактора в одной модели

Второй способ представляется более прогрессивным, т. к в этом случае появляется возможность оценить статистическую значимость влияния данного фактора на поведение эндогенной переменной на фоне других факторов, внесенных в спецификацию модели

Фиктивные переменные

Пример. Изучается зависимость расходов на образование «С» в «обычных» и «специализированных» школах в зависимости от числа учащихся N

Предположим:

1. Зависимость затрат на обучение от количества учащихся N в обоих типах школ одинакова

2. Разница в затратах объясняется необходимостью приобретения специализированного оборудования для обучения специальным дисциплинам

Тогда если строить различные модели для каждого типа школ, то спецификацию моделей можно записать в виде:

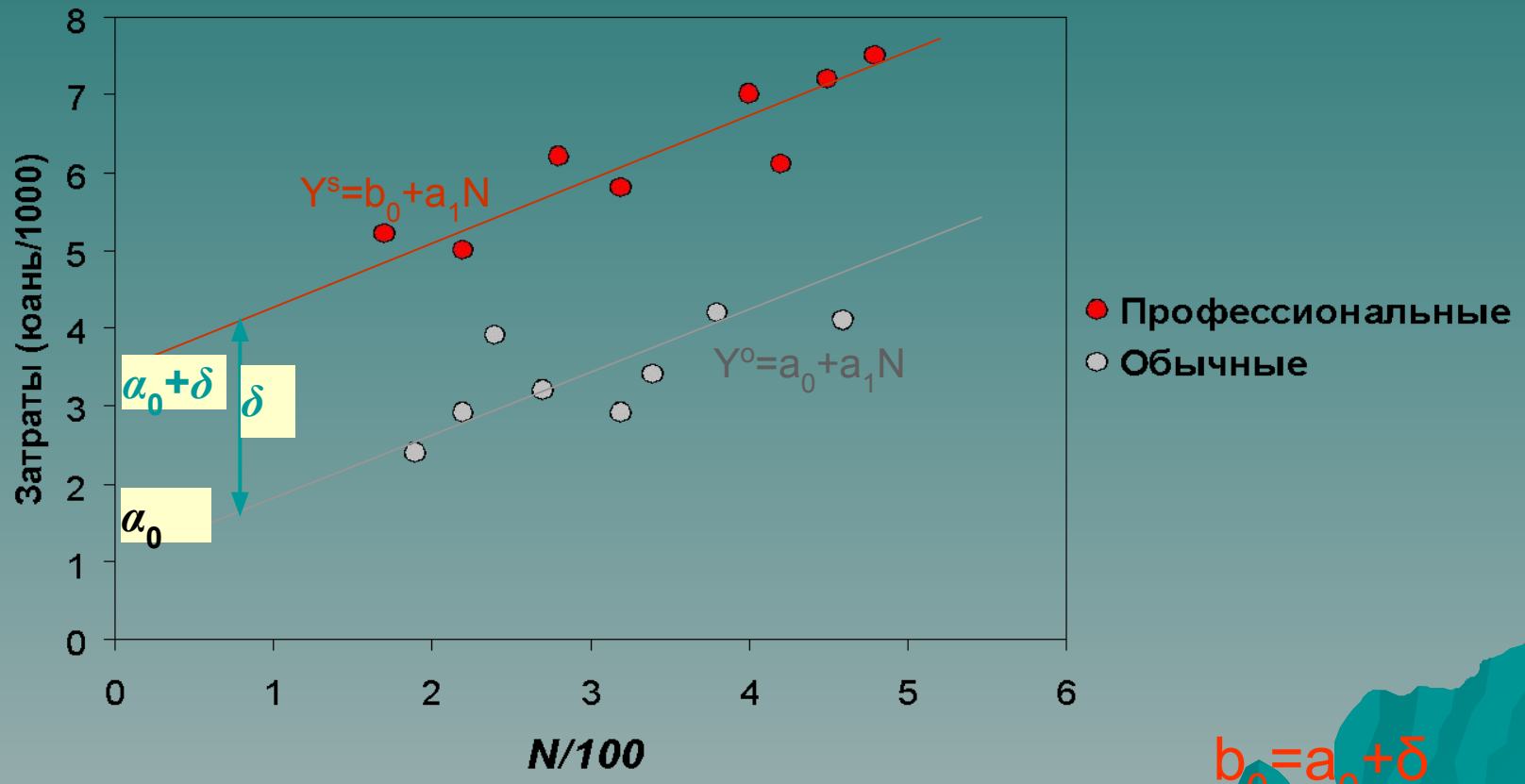
$$Y^o = a_0 + a_1 N + u$$

$$Y^s = b_0 + a_1 N + v$$

Фиктивные переменные

Пример 1 (Продолжение)

На рис.1 приведены диаграммы рассеяния и соответствующие модели для небольшой выборки школ в Китае.



$b_0=a_0+\delta$

Фиктивная переменная сдвига

Обе модели можно объединить, если ввести переменную d, область определения которой два целых числа : 0 и 1. При этом:

$$d = \begin{cases} 0 & \text{для обычных школ} \\ 1 & \text{для специализированных школ} \end{cases}$$

Спецификация такой модели имеет вид:

$$Y = a_0 + a_1 N + \delta d + u$$

Тогда при d=0 получим $Y^o = a_0 + a_1 N + u$

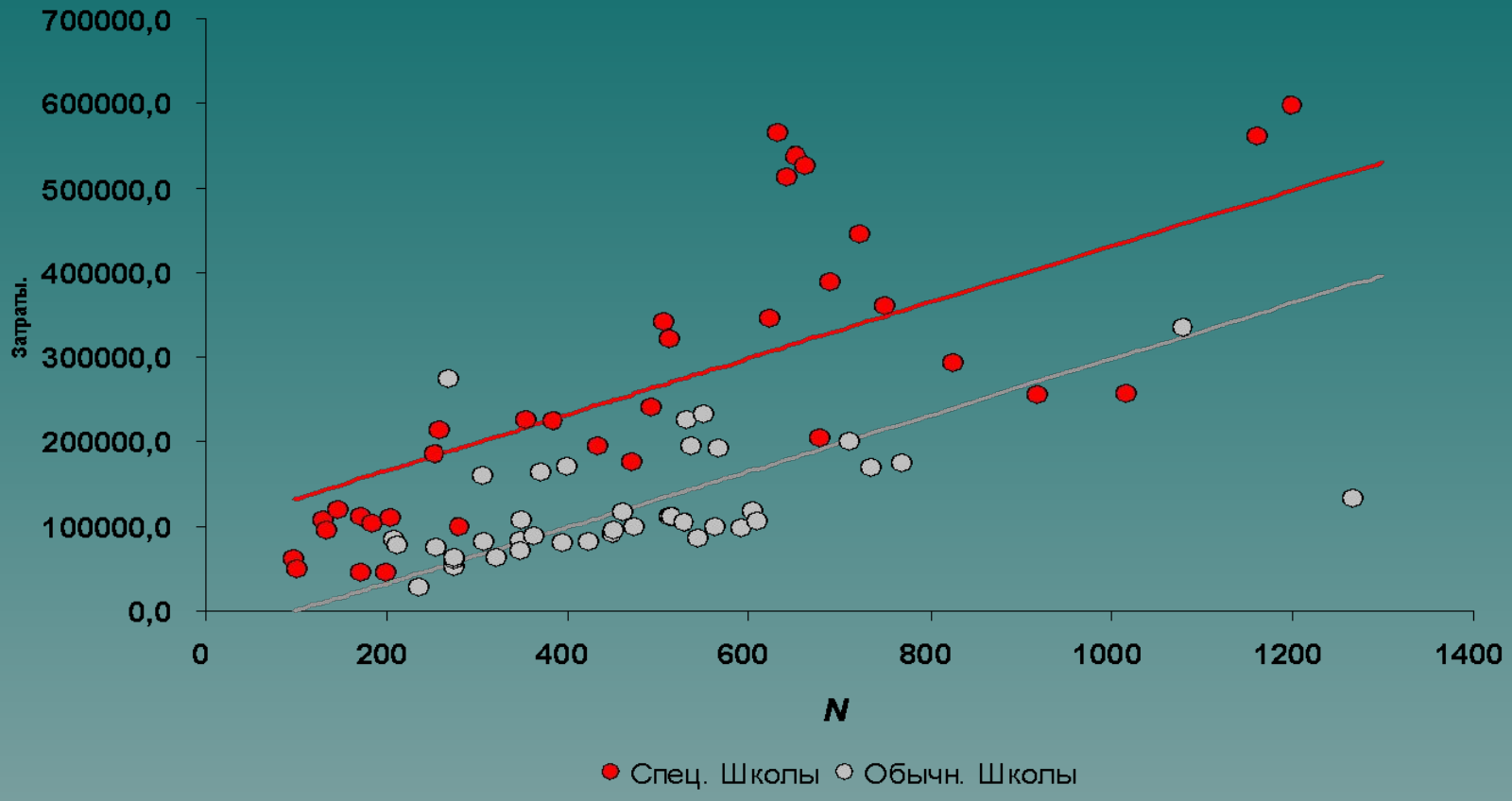
при d=1 получим $Y^s = (a_0 + \delta) + a_1 N + v$

Фиктивная переменная сдвига

Отметим:

1. Имея модель вида $Y = a_0 + a_1N + \delta d + u$, есть возможность после применения МНК оценить значения параметров a_0, a_1 и δ, стандартные ошибки их оценок, а следовательно, проверить гипотезу статистической значимости влияния фиктивной переменной d (влияние типа школ) на значения эндогенной переменной Y (затраты на обучение)
2. Графики моделей для d=0 и d=1 будут параллельны, т.к предполагается, влияние переменной N в обоих случаях остается неизменным

Фиктивная переменная сдвига



Модель Y=-33612+331.5N+133259d

соответствует Y^o = -33612 + 331.5N

Y^s= 96647 + 331.5N

Фиктивная переменная сдвига

Фиктивные переменные часто применяются при построении динамических моделей, когда с определенного момента времени начинает действовать какой-либо качественный фактор

Пример 2. Модель расходов на автотранспорт в Европе в период с 1963 по 1982 годы.

Замечание. В 1974 году в Европе начался крупный нефтяной кризис, который резко поднял цены на ГСМ.

В результате в 1974 году резко снизились расходы на автотранспорт, но затем затраты вновь стали расти с прежней скоростью.

Для учета этой ситуации вводится фиктивная переменная d, которая равна:

$$d = \begin{cases} 0 \text{ при } 1963 \le t < 1974 \\ 1 \text{ при } 1974 \le t \le 1982 \end{cases}$$

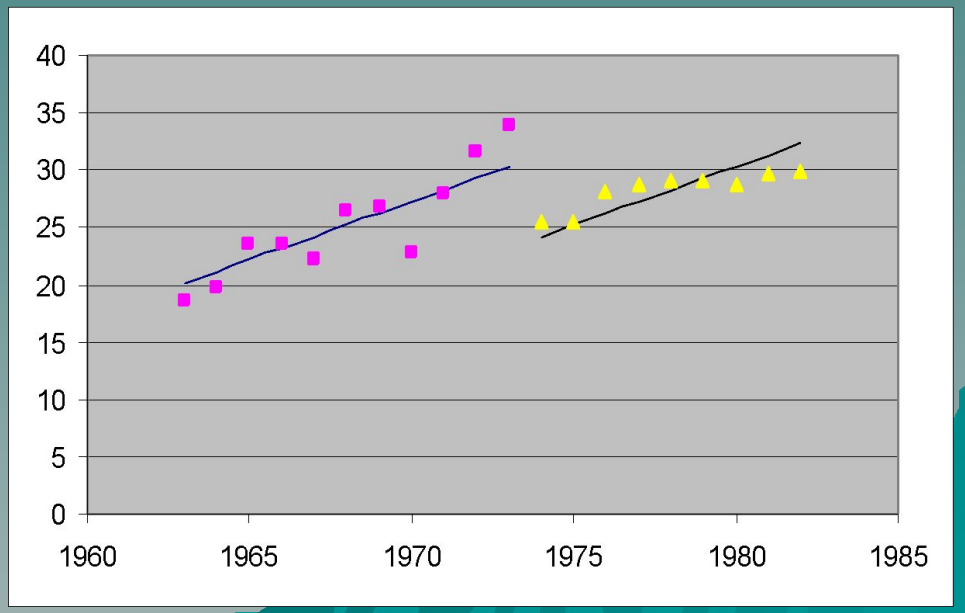
Фиктивная переменная сдвига

Год	Расходы Y	d	Время t
1963	18,5	0	0
1964	19,7	0	1
1965	23,5	0	2
1966	23,6	0	3
1967	22,2	0	4
1968	26,5	0	5
1969	26,7	0	6
1970	22,7	0	7
1971	28	0	8
1972	31,6	0	9
1973	33,9	0	10
1974	25,5	1	11
1975	25,4	1	12
1976	28,1	1	13
1977	28,8	1	14
1978	29	1	15
1979	29	1	16
1980	28,7	1	17
1981	29,6	1	18
1982	29,8	1	19

Результат ф-ции «ЛИНЕЙН»

1,0118	-7,079	20,114
0,1576	1,8268	1,0024
0,7537	2,0549	#N/A
26,016	17	#N/A
219,72	71,787	#N/A

Модель имеет вид: **Y=20.1 -7.1d +1.01t**



Фиктивная переменная сдвига (общий случай)

Пусть некоторый качественный фактор имеет несколько градаций (более 2-х)

Введение в модель фиктивных переменных с несколькими градациями рассмотрим на примере шанхайских школ, где имеются 4 категории школ: общеобразовательные, технические, ПТУ и специализированные.

Казалось достаточно ввести фиктивную переменную сдвига d, придав ей четыре различных значения и проблема будет решена.

Такой подход мало эффективен, т.к не удается оценить статистическую значимость влияния каждой градации на значения эндогенной переменной

Фиктивная переменная сдвига (общий случай)

В этом случае имеет смысл ввести отдельную переменную для каждой градации фактора.

Например:

$$d_1 = \begin{cases} 1 & \text{если школа общеобразовательная} \\ 0 & \text{если школа не общеобразовательная} \end{cases}$$

$$d_2 = \begin{cases} 1 & \text{если школа техническая} \\ 0 & \text{если школа не техническая} \end{cases}$$

$$d_3 = \begin{cases} 1 & \text{если ПТУ} \\ 0 & \text{если не ПТУ} \end{cases}$$

$$d_4 = \begin{cases} 1 & \text{если школа специализированная} \\ 0 & \text{если школа не специализированная} \end{cases}$$

Фиктивная переменная сдвига (общий случай)

Однако, если взять спецификацию модели в виде:

$$Y=a_0 + a_1d_1+a_2d_2+a_3d_3+a_4d_4+a_5N+u$$

при этом всегда верно тождество $\mathbf{d_1+d_2+d_3+d_4=1}$

Это означает, что матрица X коэффициентов системы уравнений наблюдений будет коллинеарной т.к в ней присутствует столбец из 1, и как следствие отсутствует возможность применения МНК для оценки параметров модели.

Предлагается в спецификацию ввести (к-1) фиктивную переменную (к- кол-во градаций), сделав одну из градаций базовой, относительно которой изучать влияние остальных градаций. Проблемы мультиколинеарности в этом случае не возникает

Фиктивная переменная сдвига (общий случай)

В рассматриваемом примере в качестве базового уровня можно принять градацию «Общеобразовательная»

Этой градации будет соответствовать состояние $d_2=d_3=d_4=0$

Спецификация модели примет вид:

$$\mathbf{Y=a_0+a_1N+a_2d_2+a_3d_3+a_4d_4+u} \qquad (13.1)$$

Экономический смысл коэффициентов a2, a3, a4 – превышение стоимости образования в соответствующей школе по отношению к общеобразовательной

Из уравнения (13.1) легко получить соответствующее уравнение для каждого типа школ

Фиктивная переменная сдвига (общий случай)

$Y = a_0 + a_1 N + U_1$ - Уравнение для общеобразовательных школ

$Y = (a_0 + a_2) + a_1 N + U_2$ - уравнение для «технических» школ

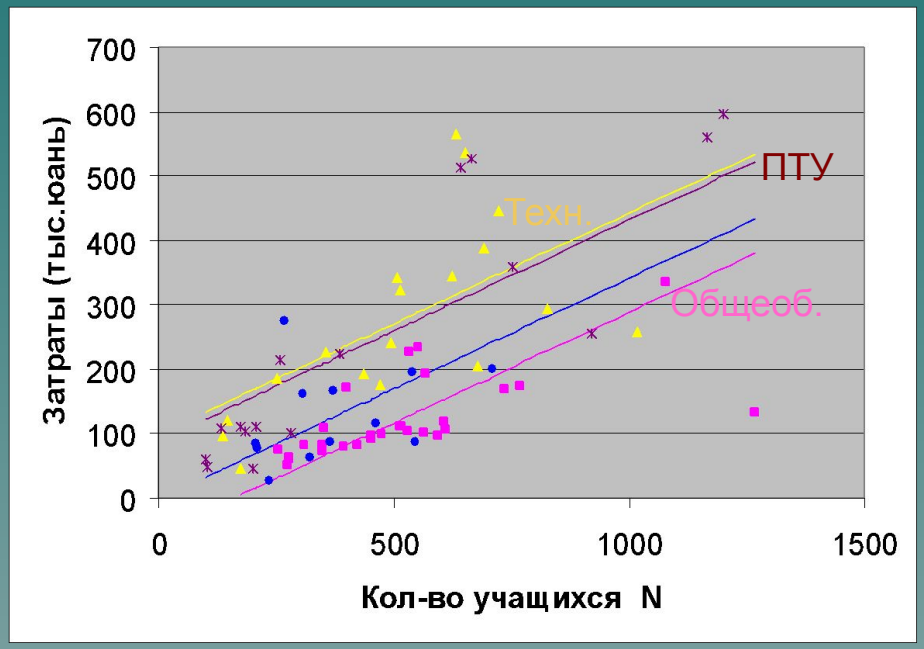
$Y = (a_0 + a_3) + a_1 N + U_3$ - уравнение для ПТУ

$Y = (a_0 + a_4) + a_1 N + U_4$ - уравнение для «специализированных» школ

Здесь также предполагается, что зависимость затрат на обучение от количества учащихся остается неизменной

Фиктивная переменная сдвига (общий случай

Результаты моделирования затрат на обучения в различных школах Шанхая



Результаты программы «ЛИНЕЙН»

53,229	143,362	154,110	0,342	-54,9
3,11	27,85	26,76	0,040	26,7
0,6	88,58	#N/A	#N/A	#N/A
29,6	69,0	#N/A	#N/A	#N/A
9,3E+08	5,4E+08	#N/A	#N/A	#N/A

Модель:

$Y = -54.9 + 0.342N + 154.11(d_2 + d_3) + 53.2d4 + U$

(26.7) (0.04) (27.9) (3.11) (88.6)

$$t = \frac{|154.11 - 143.36|}{27.85} = 0.39 < t_{кр}$$

Гипотеза H_0:(a_2=a_3) принимается

Фиктивные переменные сдвига в моделях временных рядов

Пример. Модель зависимости расходов на электроэнергию и газ в США за период 1977-1982г.г.

Год_кв.	Время t	d_2	d_3	d_4	Расходы Y
1977_1	1	0	0	0	7,33
1977_2	2	1	0	0	4,70
1977_3	3	0	1	0	5,10
1977_4	4	0	0	1	5,46
1978_1	5	0	0	0	7,65
1978_2	6	1	0	0	4,92
1978_3	7	0	1	0	5,15
1978_4	8	0	0	1	5,56
1979_1	9	0	0	0	7,96
1979_2	10	1	0	0	5,01
1979_3	11	0	1	0	5,05
1979_4	12	0	0	1	5,59

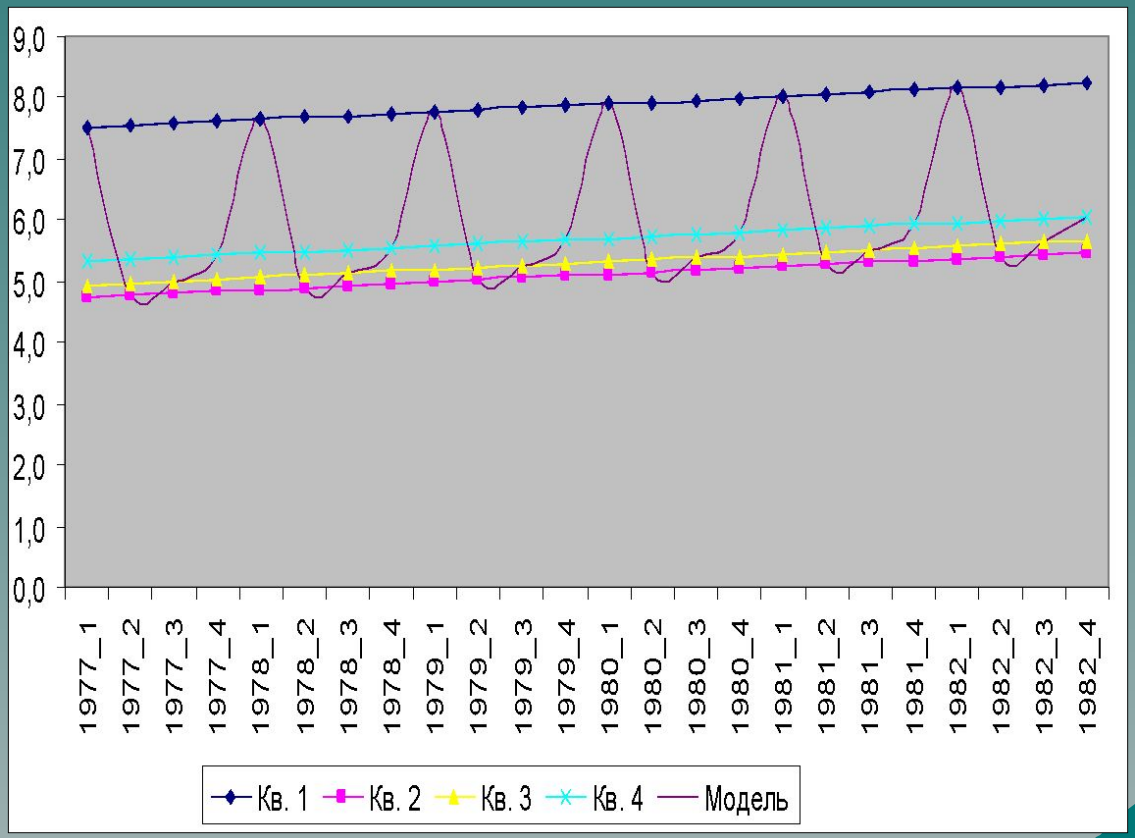
Год_кв.	Время t	d_2	d_3	d_4	Расходы Y
1980_1	13	0	0	0	7,74
1980_2	14	1	0	0	5,10
1980_3	15	0	1	0	5,67
1980_4	16	0	0	1	5,92
1981_1	17	0	0	0	8,04
1981_2	18	1	0	0	5,27
1981_3	19	0	1	0	5,51
1981_4	20	0	0	1	6,04
1982_1	21	0	0	0	8,26
1982_2	22	1	0	0	5,51
1982_3	23	0	1	0	5,41
1982_4	24	0	0	1	5,83

Фиктивные переменные сдвига в моделях временных рядов

В качестве базовой градации принят кв.1

Спецификация модели принимает вид

$$Y = a_0 + a_1 t + a_2 d_2 + a_3 d_3 + a_4 d_4 + U \qquad (13.2)$$



Результаты ф-ции «ЛИНЕЙН»

-2,19	-2,58	-2,78	0,03	7,48
0,08	0,08	0,08	0,00	0,08
0,99	0,14	#N/A	#N/A	#N/A
350,85	19,0	#N/A	#N/A	#N/A
29,47	0,40	#N/A	#N/A	#N/A

Расходы в кв.2 и кв.3 статистически не отличаются

Фиктивные переменные наклона

Во всех рассмотренных примерах априори предполагается, что различные градации качественного фактора приводят к параллельному сдвигу «базовой» модели

Это допущение не бесспорно!

В примере с различными типами школ в Шанхае предполагалось, что зависимость расходов на обучение от кол-ва учеников во всех школах одинаково

Вопрос. Как учесть эффект влияния типа школы на зависимость затрат от кол-ва учащихся?

Фиктивные переменные наклона

Для учета возможного изменения наклона графика модели при изменении градации качественного фактора предлагается ввести в спецификацию модели еще одно слагаемое вида «d умноженное на x»

Вернемся к примеру изучения зависимости расходов на образование в различных школах. Для простоты ограничимся лишь двумя градациями фактора «тип школы»: d=0 – обычная школа;

d=1 – профессиональная школа.

Спецификацию модели следует записать в виде:

$$Y = a_0 + a_1N + a_2*d + a_3dN + U \qquad (13.3)$$

Фиктивные переменные наклона

С помощью модели (13.3) появляется возможность оценить изменения наклона «базовой модели» при переходе изменении градации фактора (переменной d)

Пусть d=0, тогда модель (13.3) принимает вид:

$$\mathbf{Y= a_0 + a_1N +U_1} \qquad (13.4)$$

При d=1 получим:

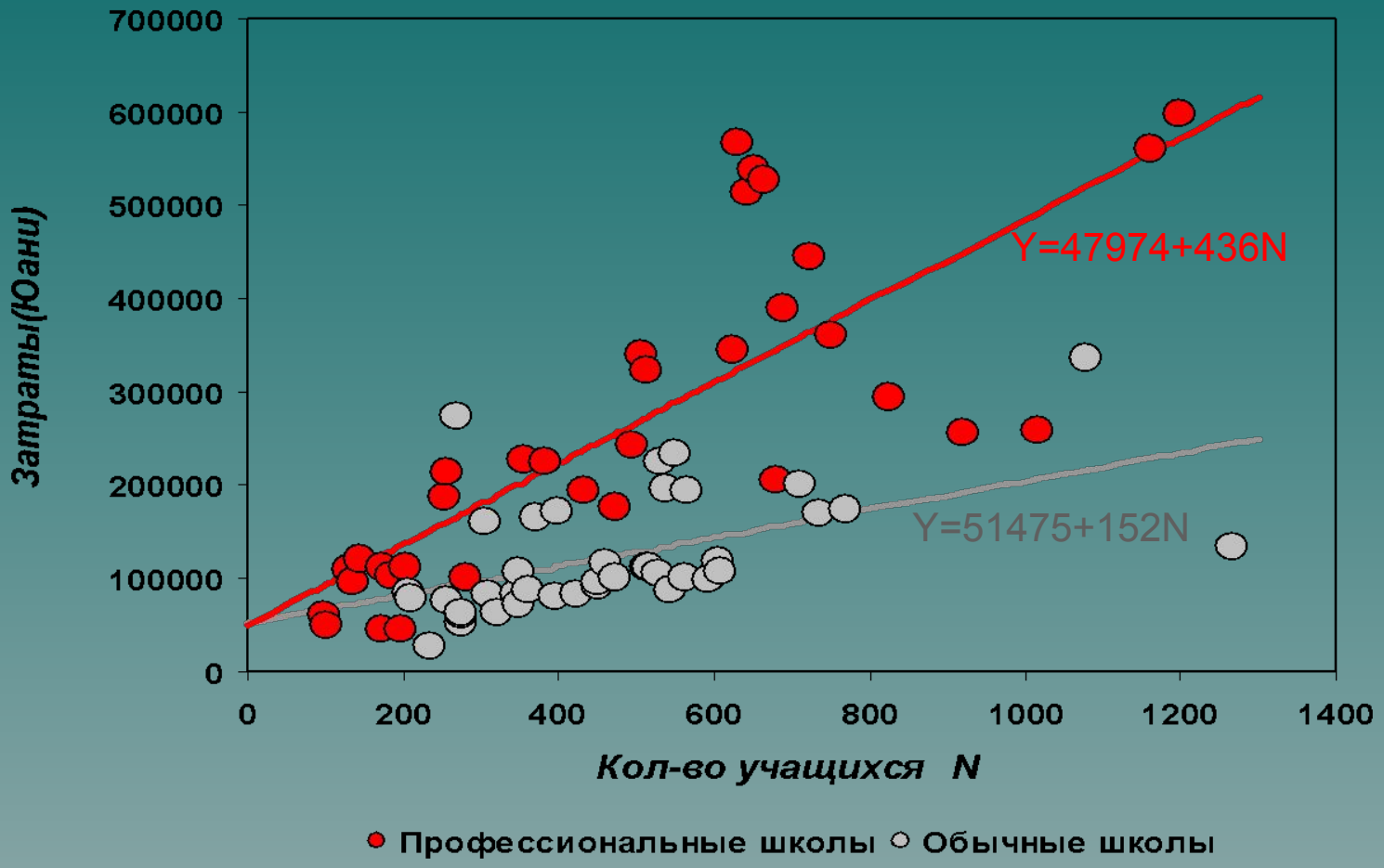
$$Y= a_0 +a_1N +a_2 +a_3N +U_2$$

или $$\mathbf{Y= (a_0+a_2) + (a_1+a_3)N +U_2} \qquad (13.5)$$

Модель (3.5), соответствующая d=1 отличается коэффициентами регрессии от модели (13.4)

В ней учитывается как «параллельный» сдвиг, так и изменение угла наклона (изменение коэффициента a_1)

Фиктивные переменные наклона



Модель: $Y=51475+152N-3501d+284dN$; $R^2=0.68$