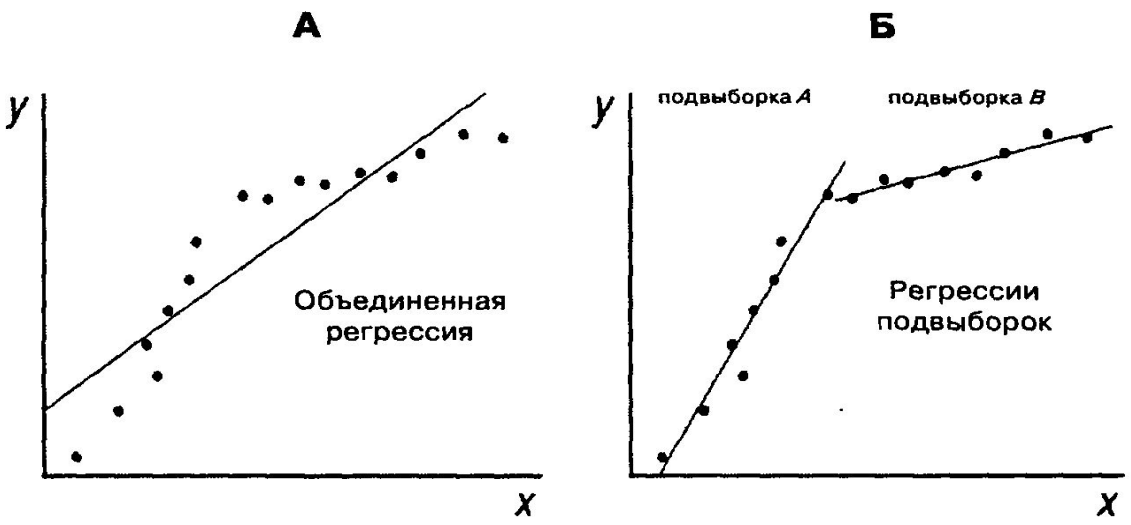
Линейные регрессионные модели с переменной структурой

Проблема неоднородных (в регрессионном смысле) данных

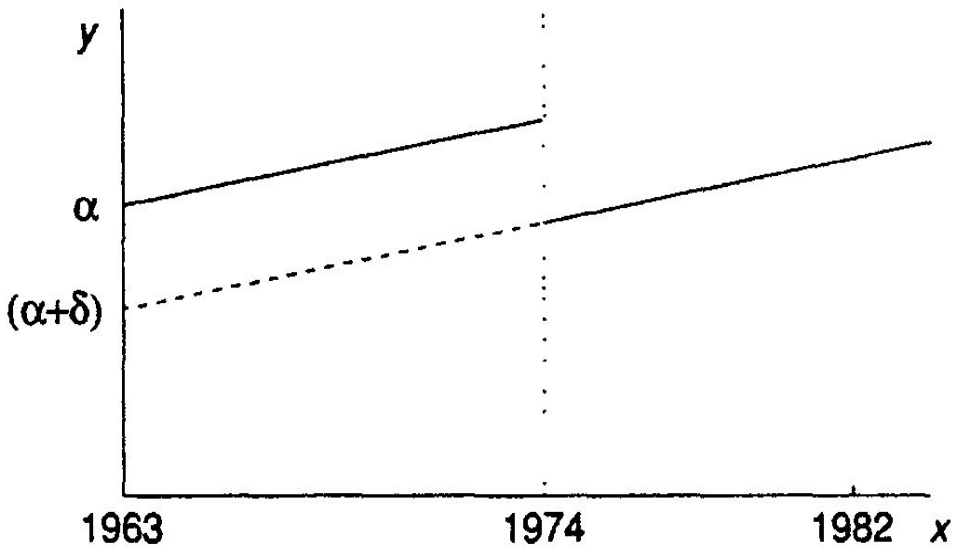
Статистические данные называются **однородными** (в регрессионном смысле), если все они зарегистрированы при одних и тех же условиях, то есть при одних и тех же значениях качественных переменных.

Статистические данные называются **неоднородными** (в регрессионном смысле), если они зарегистрированы при различных условиях (значениях качественных переменных).

Примеры неоднородных данных



Зависимость заработной платы работников (y) от производительности труда (x) с учетом пола работника



Пример статистических данных со сдвигом во времени


Качественный признак
Значения ненаблюдаемы, незарегистрированы при сборе статистических данных
Значения наблюдаемы при сборе статистических данных
Применение методов кластерного анализа (в т.ч. методы расщепления смесей вероятностных распределений) для разбиения исходных статистических данных на подвыборки объемом n', $n^{\bullet}$, …
Разбиение выборки на регрессионно однородные подвыборки объемом n', $n^{\bullet}$, …
1. Если объемы подвыборок n', n'', … значительно выше, чем $k+1$, то для каждой подвыборки оценивается своя функция регрессии
2. Объемы подвыборок n', n'', … ниже, чем $k+1$, необходимость введения фиктивных переменных, «манекенов»

Фиктивные переменные

Фиктивные переменные — бинарные переменные (принимают значения 0 или 1)

Используются для моделирования качественных признаков

Правило введения фиктивных переменных

При этом если качественная переменная имеет p градаций, то для отражения ее влияния на структуру искомой регрессионной связи необходимо ввести $(p-1)$ фиктивных переменных. Иначе для любого объекта наблюдения выполнялось бы тождество: $z_{i1} + z_{i2} + \ldots + z_{ip} = 1$, что означало бы линейную зависимость объясняющих переменных, и как следствие, невозможность получения МНК-оценок.

$$d_j^{(l)} = \begin{cases} 1, \text{ если } i-й \text{ объект обладает } l-м \textit{ качественным} \text{ свойством;} \\ 0, \textit{ в противном случае} \end{cases},$$

$$j = 1,2,\ldots, p-1$$

Фиктивные переменные: пример1

- Существует ли дискриминация на рынке труда?

$$Y_i = \beta_1 + \beta_2 * X_i + \beta_3 * D_i + \varepsilon_i$$

Y — зарплата, долларов в час

X — стаж работы, лет

$$D_i = \begin{cases} 0, \text{если } i_{\text{ый}} \text{ респондент мужчина} \\ 1, \text{если } i_{\text{ая}} \text{ респондентка женщина} \end{cases}$$

Фиктивные переменные: пример1

- Существует ли дискриминация на рынке труда?

$$Y_i = \beta_1 + \beta_2 * X_i + \beta_3 * D_i + \varepsilon_i$$

Y — зарплата, долларов в час

X — стаж работы, лет

$$D_i = \begin{cases} 0, \text{если } i_{\text{ый}} \text{ респондент мужчина} \\ 1, \text{если } i_{\text{ая}} \text{ респондентка женщина} \end{cases}$$

Фиктивные переменные: пример1

- Существует ли дискриминация на рынке труда?

$$Y_i = \beta_1 + \beta_2 * X_i + \beta_3 * D_i + \varepsilon_i$$

Y — зарплата, долларов в час

X — стаж работы, лет

$$D_i = \begin{cases} 0, \text{если } i_{\text{ый}} \text{ респондент мужчина} \\ 1, \text{если } i_{\text{ая}} \text{ респондентка женщина} \end{cases}$$

Фиктивные переменные: пример 2

Исследовать зависимость веса новорожденного (y) от среднего числа сигарет (x), выкуриваемых матерью в день, с учетом уже имеющихся у матерей детей z

	1 Y	2 X	3 Z
1	3,52	10	1
2	3,46	19	2
3	3,2	16	0
4	3,32	28	1
5	3,54	4	3
6	3,31	14	2
7	3,36	21	0
8	3,65	10	1
9	3,15	22	0
10	3,44	12	1
11	3,1	31	3
12	3,22	29	0
13	3,71	8	2
14	3,76	6	1
15	3,92	8	1

Фиктивные переменные: пример 2

1. Построение регрессионной модели для отдельных подвыборок матерей по количеству уже имеющихся детей, приводит к невозможности проведения регрессионного анализа, поскольку некоторые группы (3 и 4) содержат небольшое количество наблюдений

Количество детей у матери (z)	0	1	2	3
Количество матерей	4	6	3	2

2. Для учета наличия структурных сдвигов, введем фиктивные переменные

$$d_i = \begin{cases} 1 - & \text{если кол-во имеющихся у матери детей } i \\ 0 - & \text{в противном случае} \end{cases}$$

Таблица

Y	X	Z	d1	d2	d3
3,52	10	1	1	0	0
3,46	19	2	0	1	0
3,2	16	0	0	0	0
3,32	28	1	1	0	0
3,54	4	3	0	0	1
3,31	14	2	0	1	0
3,36	21	0	0	0	0
3,65	10	1	1	0	0
3,15	22	0	0	0	0
3,44	12	1	1	0	0
3,1	31	3	0	0	1
3,22	29	0	0	0	0
3,71	8	2	0	1	0
3,76	6	1	1	0	0
3,92	8	1	1	0	0

Фиктивные переменные: пример 2

N=15	Regression Summary for Dependent Variable: Y (Spreadsheet1) R= ,86521807 R?= ,74860231 Adjusted R?= ,68003930 F(2,12)=10,918 p<,00126 Std.Error of estimate: ,13604 Beta	Std.Err. of Beta	B	Std.Err. of B	t(11)	p-level
Intercept			3,612701	0,110439	32,71222	0,000000
X	-0,622960	0,169405	-0,017124	0,004657	-3,67735	0,003643
d1	0,422031	0,183996	0,200157	0,087264	2,29370	0,042500

$$\hat{y} = \underset{(0,110)}{3,61} - \underset{(0,005)}{0,017}\,x + \underset{(0,087)}{0,200}\,d_1$$

Вывод: **1.** при рождении вес второго ребенка у курящих женщин будет в среднем на 0,2 кг выше, чем вес детей у остальных женщин, выкуривающих в день то же количество сигарет;

2. с ростом количества выкуриваемых сигарет вес новорожденного будет уменьшаться на 0,017 кг.

Критерий ЧОУ

$$H_0 : \overline{\beta}^{(1)} = \overline{\beta}^{(2)}; \sigma_\varepsilon^2(1) = \sigma_\varepsilon^2(2)$$

$$H_1 : \overline{\beta}^{(1)} \neq \overline{\beta}^{(2)}; \sigma_\varepsilon^2(1) \neq \sigma_\varepsilon^2(2)$$

Для проверки гипотезы используется статистика Чоу, которая в условиях справедливости нулевой гипотезы распределена по закону Фишера-Снедекора $(\nu_1 = n_2, \nu_2 = n_1 - k - 1)$:

$$\gamma_{n,n_1} = \frac{\left(e^T e - e^{(1)T} e^{(1)}\right)/n_2}{e^{(1)T} e^{(1)}/(n_1 - k - 1)},$$

где e - вектор регрессионных остатков, оцененных по всей выборке;

$e^{(1)}$ - вектор регрессионных остатков, оцененных по первой подвыборке.

Если n_2 достаточно велико, то наряду с данным подходом используется другой критерий, связанный с критерием Чоу:

$$\gamma_{n_1,n_{21}} = \frac{\left(e^T e - e^{(1)T} e^{(1)} - e^{(2)T} e^{(2)}\right)/(k+1)}{\left(e^{(1)T} e^{(1)} + e^{(2)T} e^{(2)}\right)/(n_1 + n_2 - 2k - 2)},$$