

# Лекция 6

## Модели с дискретными переменными

- 1. Фиктивные объясняющие переменные*
- 2. Модели с дискретными зависимыми переменными*
- 3. Тесты Гуйарати и Чоу.*

# 1. Фиктивные объясняющие переменные

До сих пор рассматривались модели, в которых в качестве объясняющих переменных выступали *количественные* переменные, т.е. признаки, принимающие любые значения из некоторого числового множества (доход семьи, производительность, себестоимость и т.д.).

На практике возникает необходимость исследования влияния на зависимую переменную *качественных* признаков, которые могут принимать два или более фиксированных уровней, не являющихся числовыми, а являющимися некоторыми категориями.

Примерами таких признаков могут служить: образование (начальное, среднее, высшее), пол человека (мужской, женский) и т.д.

Чтобы учесть такие признаки в модели, они должны быть преобразованы в количественные, т.е. им должны быть присвоены *количественные метки*. Сконструированные на основе качественных факторов числовые переменные называют *фиктивными переменными* (двоичными, индикаторными).

Такие переменные приводят к скачкообразному изменению параметров регрессионных моделей и в этом случае говорят об исследовании моделей с *переменной структурой*.

Регрессионные модели, содержащие лишь качественные факторы, называются *ANOVA – моделями* (моделями дисперсионного анализа). Например, зависимость заработной платы от образования может быть представлена в виде:

$$y_i = \alpha_0 + \alpha_1 z_i + \varepsilon_i,$$

где ,  $z_i = 0$  если  $i$  –й персоналий не имеет высшего образования и  $z_i = 1$  в противном случае.

Нетрудно видеть, что ANOVA – модели представляют собой кусочно-постоянные функции, и они достаточно редко используются в экономике.

Чаще встречаются модели, содержащие как количественные, так и качественные факторы.

Такие модели называют *ANCOVA-моделями* (модели ковариационного анализа).

Обычно в качестве фиктивных переменных выступают бинарные переменные, т.е. переменные, принимающие только два значения: 0 и 1. Например, заработная плата го служащего предприятия может быть представлена следующей моделью:

$$y_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ji} + \alpha_1 z_{1i} + \varepsilon_i, \quad i = \overline{1, n},$$

где  $z_{1i} = 1$ , если  $i$  – служащий является мужчиной, и  $z_{1i} = 0$ , если  $i$  – служащий является женщиной,  $x_{ji}$  – количественные признаки (стаж работы, возраст и т.д.),  $n$  – число служащих предприятия.

Коэффициент  $\alpha_1$  в этой модели называют *дифференциальным* свободным членом, ибо он показывает, на какую величину изменится свободный член модели при изменении переменной  $z_{1i}$ .

Если рассматриваемый качественный признак имеет более чем два уровня, например, их число равно  $k$  ( $k > 2$ ), то в рассмотрение вводят  $(k - 1)$  бинарную фиктивную переменную.

В рассматриваемом примере о заработной плате для учета влияния фактора образования (начальное, среднее, высшее, т.е.  $k = 3$ ) на величину заработной платы необходимо ввести дополнительно в модель  $k - 1 = 2$  бинарные переменные  $Z_{2i}$  и  $Z_{3i}$ :

$$y_i = \beta_0 + \sum \beta_j x_{ji} + \alpha_1 z_{1i} + \alpha_2 z_{2i} + \alpha_3 z_{3i} + \varepsilon_i, \quad i = \overline{1, n} \quad (1)$$

В данной модели

$$z_{2i} = \begin{cases} 1, & \text{если } i\text{-й служ. имеет ВО;} \\ 0, & \text{в других случаях,} \end{cases}$$

$$z_{3i} = \begin{cases} 1, & \text{если } i\text{-й служ имеет СО;} \\ 0, & \text{в других случаях.} \end{cases}$$

Как видим, третьей фиктивной переменной не требуется, так как при  $z_{2i} = z_{3i} = 0$  следует, что  $i$  – служащий имеет начальное образование.

Нулевой уровень фиктивных переменных называется *базовым* или *сравнительным уровнем* модели.

Оценку коэффициентов модели (1) в том числе и при фиктивных переменных выполняют МНК по той же схеме, как и при количественных факторах модели, описанной выше.

## 2. Модели с дискретными зависимыми переменными

Нередко зависимая переменная по своей природе является *дискретной*, например, если исследовать зависимость количество автомобилей в семье от уровня доходности и других факторов, то видно, что эта переменная принимает целые значения: 0, 1, 2, ... .

Изучим несколько типичных ситуаций и выделим основные виды таких переменных.

# *Номинальные переменные.*

Рассмотрим следующие примеры.

1. Семейное положение мужчины можно выразить следующими категориями: холост, женат, разведен, вдовец.
2. Решение о покупке товара: да, нет.
3. Выбор специальности при поступлении в институт: коммерсант, менеджер, экономист.

Выбор значения осуществляется из двух или более альтернатив.

Если имеется только две возможности, то наблюдения обычно описываются бинарной переменной.

В общем случае при наличии  $k$  альтернатив результат можно описать переменной, принимающей только целые значения:  $1, 2, 3, \dots, k$ .

Главная особенность приведённых примеров состоит в том, что имеющиеся альтернативы нельзя *естественным образом упорядочить*, их нумерация от 1 до  $k$  может быть произвольной и зависит от исследователя. Такие переменные называют *номинальными*.

### *Порядковые переменные.*

Как и в предыдущем случае имеется несколько альтернатив, но они могут быть естественным образом упорядочены.

В качестве примеров рассмотрим:

1. Доход семьи: низкий, средний, высокий, очень высокий.
2. Уровень образования: начальное, незаконченное среднее, среднее, незаконченное высшее, высшее.
3. Состояние больного: плохое, удовлетворительное, хорошее.

Такие переменные называют *порядковыми* или *ранговыми*.

## *Количественные целочисленные переменные.*

Примерами таких переменных служат:

1. Число предприятий страны, обанкротившихся в текущем году.
2. Количество частных вузов в городе.
3. Число прибыльных фирм города

Для моделей с описанными дискретными зависимыми переменными возможно формальное применение МНК для оценки их коэффициентов.

Однако с содержательной точки зрения удовлетворительные результаты можно получить только для моделей с количественными целочисленными переменными.

Если зависимая переменная является *номинальной* и количество альтернатив более двух, то результаты оценивания МНК вообще теряют смысл в силу произвольной нумерации альтернатив.

Поэтому стандартная схема оценки параметров модели в случае номинальных зависимых переменных нуждается в существенной коррекции.

Рассмотрим вначале простейшие модели *бинарного выбора*, когда результирующий показатель может принимать только два значения: 0 и 1.

Изучим свойства таких моделей на примере покупки некоторой  $i$  – й семьёй автомобиля. Будем считать  $y_i = 1$ , если в течение исследуемого периода семья приобретёт автомобиль и  $y_i = 0$  – в противном случае.

На решение о покупке автомобиля влияют различные факторы: доход семьи, количество членов семьи, их возраст, место проживания и т.д. Набор этих факторов можно представить вектором  $x = (x_1, x_2, \dots, x_p)$ .

На решение семьи влияют также неучтенные и случайные (расходы на лечение случайной болезни, расходы на ремонт квартиры после затопления соседями и т.д.) факторы  $\varepsilon$ .

Выдвигая различные предположения о характере зависимости переменной  $Y$  от вектора  $X$  и случайного фактора  $\varepsilon$ , можно получить различные модели бинарного выбора.

Например, можно воспользоваться обычной линейной моделью регрессии:

$$y_i = \sum_{j=1}^p \beta_j x_{ji} + \varepsilon_i, \quad i = \overline{1, n}. \quad (2)$$

Поскольку  $y_i$ , как случайная величина, принимает только два значения (0 и 1), а по предпосылке 2° МНК верно равенство

$$M(\varepsilon_i) = 0,$$

то, находя математическое ожидание зависимой переменной, получим с учетом предпосылки 1°:

$$\begin{aligned} M(y_i) &= 1 \cdot P(y_i = 1) + 0 \cdot P(y_i = 0) = P(y_i = 1) = \\ &= M\left(\sum_{j=1}^p \beta_j x_{ji} + \varepsilon_i\right) = \sum_{j=1}^p \beta_j x_{ji}. \end{aligned}$$

В итоге модель (2) может быть записана в следующем виде

$$P(y_i = 1) = \sum_{j=1}^p \beta_j x_{ji}. \quad (3)$$

и поэтому её называют *линейной моделью вероятности*.

Нетрудно показать, что модель (3) является гетероскедастичной. Другим важным недостатком модели является тот факт, что прогнозное значение зависимой переменной, вычисленное по полученному выборочному уравнению регрессии (правая часть уравнения (3))

$$\tilde{y}_i = b_1 x_{1i} + b_2 x_{2i} + \dots + b_p x_{pi}$$

может находиться вне отрезка  $[0,1]$  , что не поддается разумной интерпретации, поскольку левая часть уравнения (3) представляет вероятность.

От указанного недостатка, связанного с предположением о линейной зависимости вероятности  $P(y_i = 1)$  от вектора  $x$ , можно избавиться, если предположить что данная зависимость является нелинейной

$$P(y_i = 1) = F(\beta_1, \beta_2, \dots, \beta_p, x_{1i}, x_{2i}, \dots, x_{pi}), \quad (4)$$

где  $F(\dots)$  – некоторая функция с областью значений на отрезке  $[0,1]$ .

В частности, в качестве можно взять функцию распределения вероятностей некоторой случайной величины.

Наиболее распространенными функциями такого вида являются:

1. В качестве  $F$  рассматривается функция стандартного нормального распределения вероятностей

$$F(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^u e^{-\frac{x^2}{2}} dx,$$

и в этом случае модель (4) называют *probit-моделью*.

2. Если в качестве  $F$  выбирают логистическую функцию

$$F(u) = \frac{e^u}{1 + e^u} = \frac{1}{1 + e^{-u}}$$

то говорят о *logit-модели*.

Для оценивания коэффициентов probit- и logit-моделей обычно используют *метод максимального правдоподобия*.

В том случае, когда номинальная зависимая переменная  $y$  имеет более двух альтернатив, т.е. требуется построить *модель множественного выбора*, то используют различные подходы. Один из них заключается в представлении модели как последовательности бинарных выборов.

Допустим, что изучается выбор одной из трёх профессий: инженера, экономиста, юриста. Вводят в рассмотрение две бинарные переменные:

$$y_u = \begin{cases} 1 & \text{для инженера,} \\ 0 & \text{для других профессий,} \end{cases}$$

$$y_{\text{эк}} = \begin{cases} 1 & \text{для экономиста,} \\ 0 & \text{для иных профессий.} \end{cases}$$

Тогда выбор одного из трёх вариантов профессий можно описать в виде графа последовательных действий, в вершинах которого происходит бинарный выбор (рис. 1).

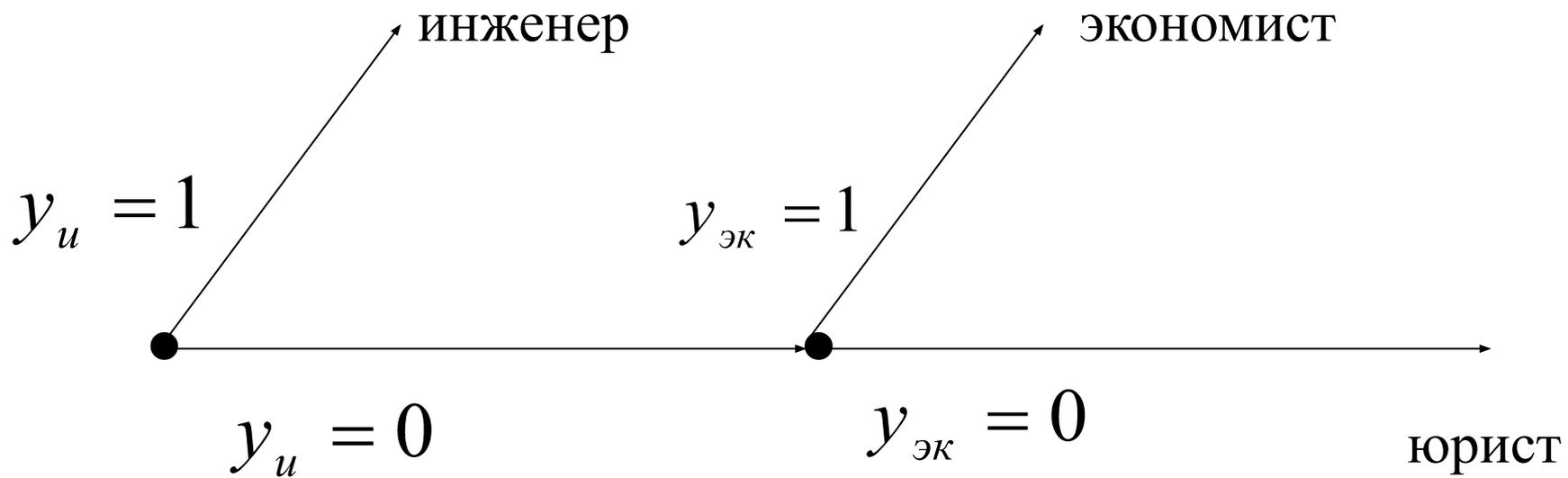


Рис. 1

### 3. Тесты Гуйарати и Чоу

Пусть требуется оценить парную регрессию, в которой в качестве объясняющей переменной выступает время :

$$y_i = \beta_0 + \beta_1 t_i + \varepsilon_i, \quad i = \overline{1, n}$$

Предположим, что в момент времени  $t^*$  произошло изменение характера динамики изучаемого показателя , вызванные структурными изменениями в экономике (экономический кризис, природные катаклизмы и т. д.).

Пусть до момента  $t^*$  было произведено  $n_1$  наблюдений показателя  $y$ , а после этого момента -  $n_2$ . В итоге в сумме  $n_1 + n_2 = n$ .

Тогда одной из задач анализа процесса является выяснения вопроса о том, значительно повлияли общие структурные изменения на параметры модели. Если это влияние значительно, то для моделирования зависимости  $y$  от времени  $t$  следует использовать *кусочно-линейные* модели регрессии, т.е. одна модель будет описывать процесс до момента времени  $t^*$ , а другая – после него.

Если же структурные изменения незначительно повлияли на характер динамики, то её описывают единым по всей совокупности уравнением регрессии.

Для ответа на этот вопрос в тесте **Гуй-арати** в модель регрессии включается фиктивная переменная  $z$  :

$$y_i = \beta_0 + \beta_1 z_i + \beta_2 t_i + \beta_3 (z_i \cdot t_i) + \varepsilon_i, \quad (5)$$

где

$$z_i = \begin{cases} 1 & \text{для } t_i > t^*, \\ 0 & \text{для } t_i \leq t^*. \end{cases}$$

В итоге для каждого промежутка времени получаются следующие оценки уравнения регрессии:

- для  $t_i \leq t^*$ :  $\tilde{y}_i = b_0 + b_2 t_i$  ;

- для  $t_i > t^*$ :  $\tilde{y}_i = (b_0 + b_1) + (b_2 + b_3) t_i$  .

С помощью  $t$  – критерия Стьюдента проверяют значимость полученных оценок  $b_j$  коэффициентов регрессии  $\beta_j$  (5).

Здесь возможны следующие случаи.

1°. Если  $b_1$  статистически значим, а параметр  $b_3$  – нет, то изменение динамики вызвано различием свободных членов регрессии кусочно-линейной модели (рис. 2).

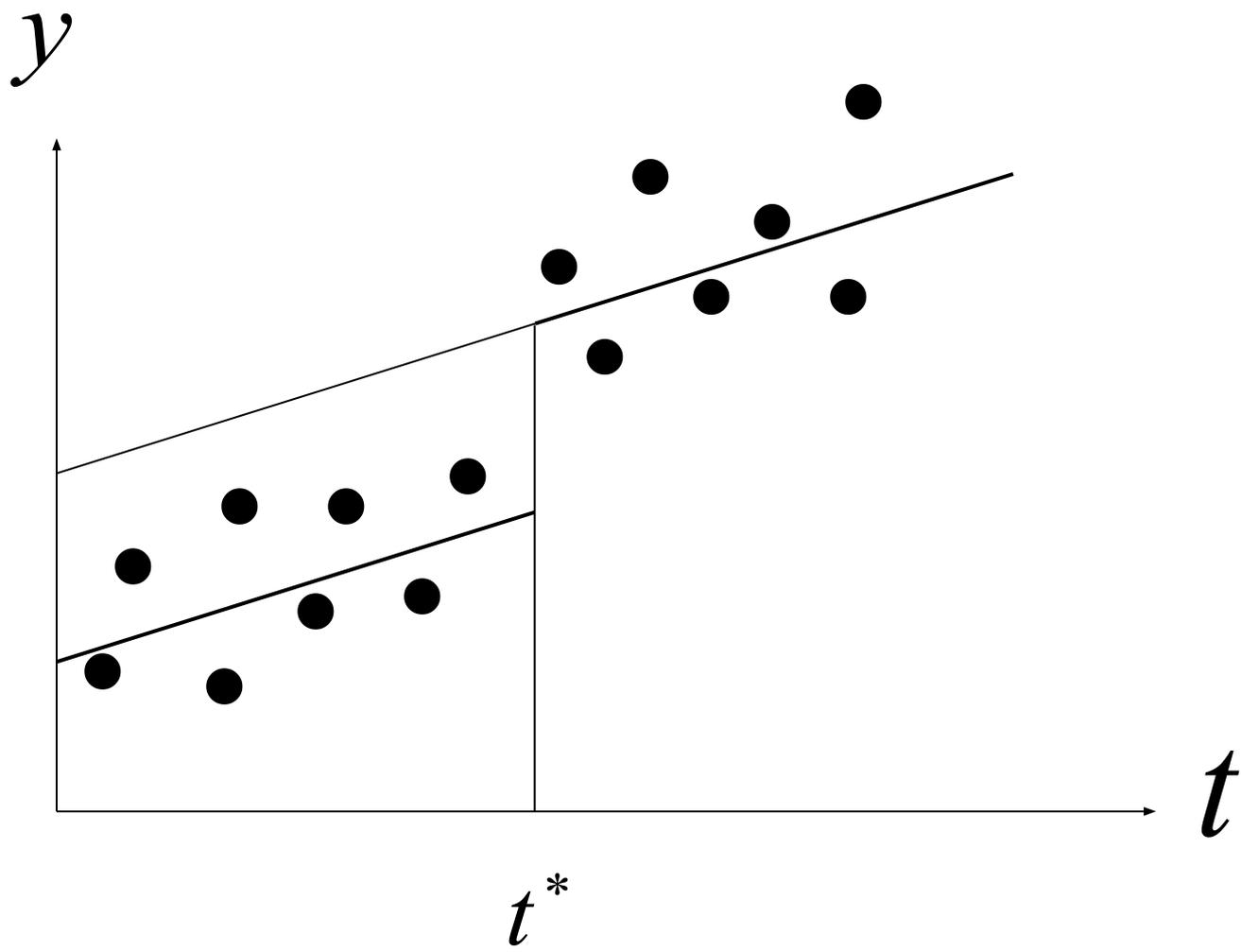


Рис. 2

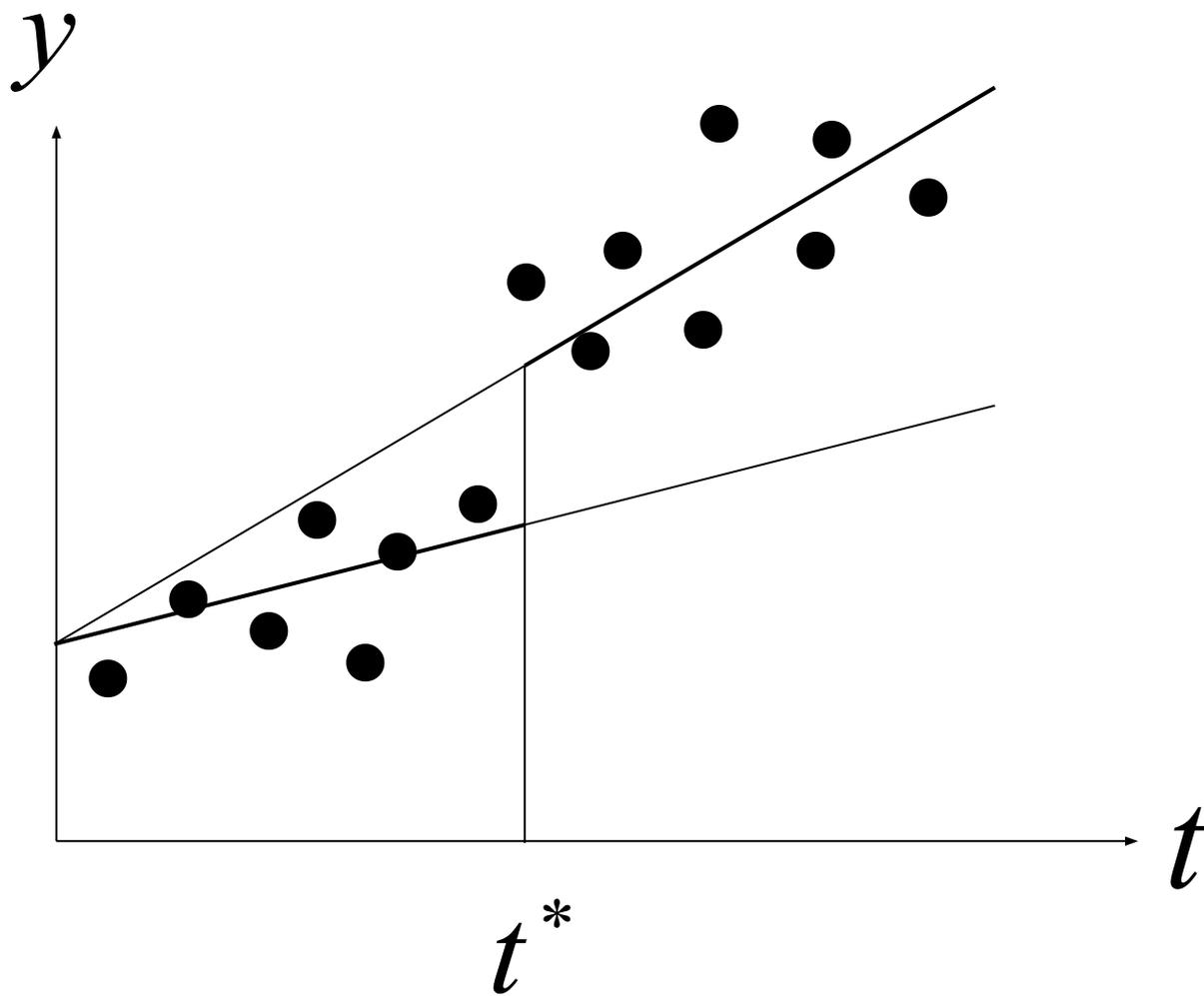


Рис. 3

2°. Если параметр  $b_3$  статистически значим, а  $b_1$  не является значимым, то различаются коэффициенты регрессии кусочно-линейной модели (рис. 3).

3°. Если оба параметра  $b_1$  и  $b_3$  статистически значимы, то изменение зависимости признака  $y$  от времени  $t$  вызвано как различием свободных членов, так и коэффициентов регрессии (рис. 4).

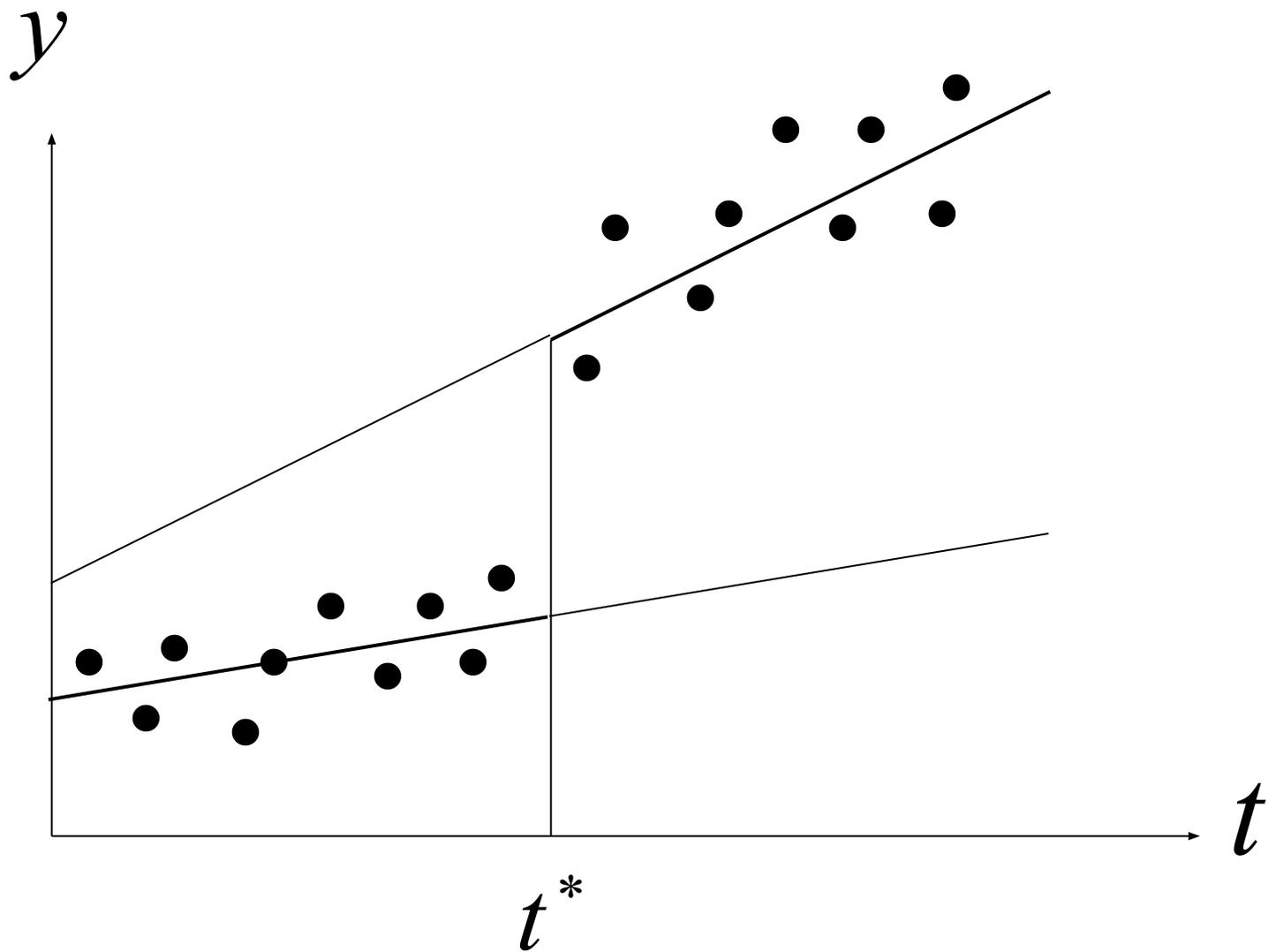


Рис. 4

4°. Если оба параметра  $b_1$  и  $b_3$  статистически незначимы, то используется *единая* по всей совокупности данных линейная регрессия, т. е. структурные изменения в экономике незначительно повлияли на характер динамики переменной  $y$ .

Целесообразность применения двух уравнений регрессии вместо одного можно оценить, не прибегая к фиктивным переменным. Для этого используют тест Г. Чоу.

Выдвигается гипотеза  $H_0$  о незначительном влиянии структурных изменений в экономике. Согласно тесту Чоу гипотеза  $H_0$  отвергается на уровне значимости  $\alpha$  (т.е. требуется кусочно-линейная модель), если

статистика

$$F = \frac{\sum_{i=1}^n e_i^2 - \left( \sum_{i=1}^{n_1} e_i^2 + \sum_{i=n_1+1}^n e_i^2 \right)}{\sum_{i=1}^{n_1} e_i^2 + \sum_{i=n_1+1}^n e_i^2} \cdot \frac{n - p_1 - p_2 - 2}{p_1 + p_2 - p_3 + 1} \quad (6)$$

больше  $F_{кр}$ , найденного по таблицам по заданному уровню значимости  $\alpha$  и числу степеней свободы

$$k_1 = p_1 + p_2 - p_3 + 1, \quad k_2 = n - p_1 - p_2 - 2.$$

В формуле (6)  $p_1, p_2, p_3$  — число параметров (без свободного члена) в уравнениях, построенных по статистическим данным до времени  $t^*$ , после него и по всей совокупности данных соответственно.

Таким образом, в тесте Чоу в отличие от теста Гуйарати требуется построить три уравнения регрессии:

● по всей выборке (чтобы найти  $\sum_{i=1}^n e_i^2$ );  
● по выборке до времени  $t^*$  (чтобы

определить  $\sum_{i=1}^{n_1} e_i^2$ );

● по выборке после  $t^*$  (чтобы вычислить

$\sum_{i=n_1+1}^n e_i^2$ ).