Фиктивные Переменные

1. Типы фиктивных переменных.
2. Тест Чоу

Фиктивная переменная (ФП) – это переменная, которая принимает два различных значения.

Эти различные значения могут быть любыми числами, но в целях удобства интерпретации это всегда

0 и 1.

ФП используются для ввода в модель регрессии качественных и категориальных факторов.

I. ФП для качественного фактора, принимающего два значения.

Модель без взаимодействия.

На фактор **Y**, кроме количественных факторов X_2, X_3, …, X_k, воздействует качественный фактор, который принимает два значения (имеет две категории):

А и Б,

или

А и не А.

Чтобы учесть влияние этого фактора, в модель вводят фиктивный фактор D.

$$D = \begin{cases} 0 \\ 1 \end{cases}$$

для объектов, на которых качественный фактор принимает значение А

для объектов, на которых качественный фактор принимает значение не А

Или можно наоборот:

$$D = \begin{cases} 0 & \text{для …не A} \\ 1 & \text{для … A} \end{cases}$$

Модель тогда имеет вид:

$$Y = \beta_1 + \beta_2 * X_2 + \ldots + \beta_k * X_k + \delta * D + u$$

$$Y = \beta_1 + \beta_2 * X_2 + \ldots + \beta_k * X_k + \delta * D + u$$

Интерпретация коэффициента **δ**:

при любых фиксированных значениях факторов $x_2, x_3, \ldots, x_k$ значения фактора **Y** различаются в среднем на **δ** для объектов, на которых качественный признак **D** принимает и не принимает значение A.

$$Y = \beta_1 + \beta_2 * X_2 + \ldots + \beta_k * X_k + \delta * D + u$$

Проверяя по t-тесту значимость **δ**, мы тем самым проверяем значимость или незначимость различия значений **Y** для объектов имеющих и не имеющих качество А.

ПРИМЕР 1.

Y – среднемесячное потребление семьи, в рублях.

X – среднемесячный доход семьи, в рублях.

Предполагается, что потребление зависит также от того, проживает ли семья в городе или в сельской местности.

Вводим ФП D. Пусть D=1 для семей из сельской местности и D=0 для городских семей.

Модель:

$$Y = \beta_1 + \beta_2 * X + \delta * D + u.$$

Модель оценивается по выборке n=30.

$$\hat{Y} = 3750 + 0{,}57*X - 1230*D$$

(1119) (0.22) (349)

Проверяем гипотезу:

$H_0: \delta = 0$

$H_A: \delta \neq 0$

Гипотеза H_0 отвергается при у.з. 1%.

Вывод: существует значимое различие в затратах на потребления для городских и сельских семей, имеющих одинаковый доход.

Сельские семьи тратят на потребление в среднем на 1230 рублей меньше, чем городские семьи, имеющие такой же доход.

Замечание: в теоретической модели предполагается, что на изменение дохода городские и сельские семьи реагируют одинаково.

При каждом увеличении дохода на 1 руб. потребление обоих типов семей увеличивается в среднем на 0,57 рубля.

$$\hat{Y} = 3750 + 0{,}57*X - 1230*D$$

Можно получить уравнения отдельно для сельских и городских семей.

Для городских D=0:

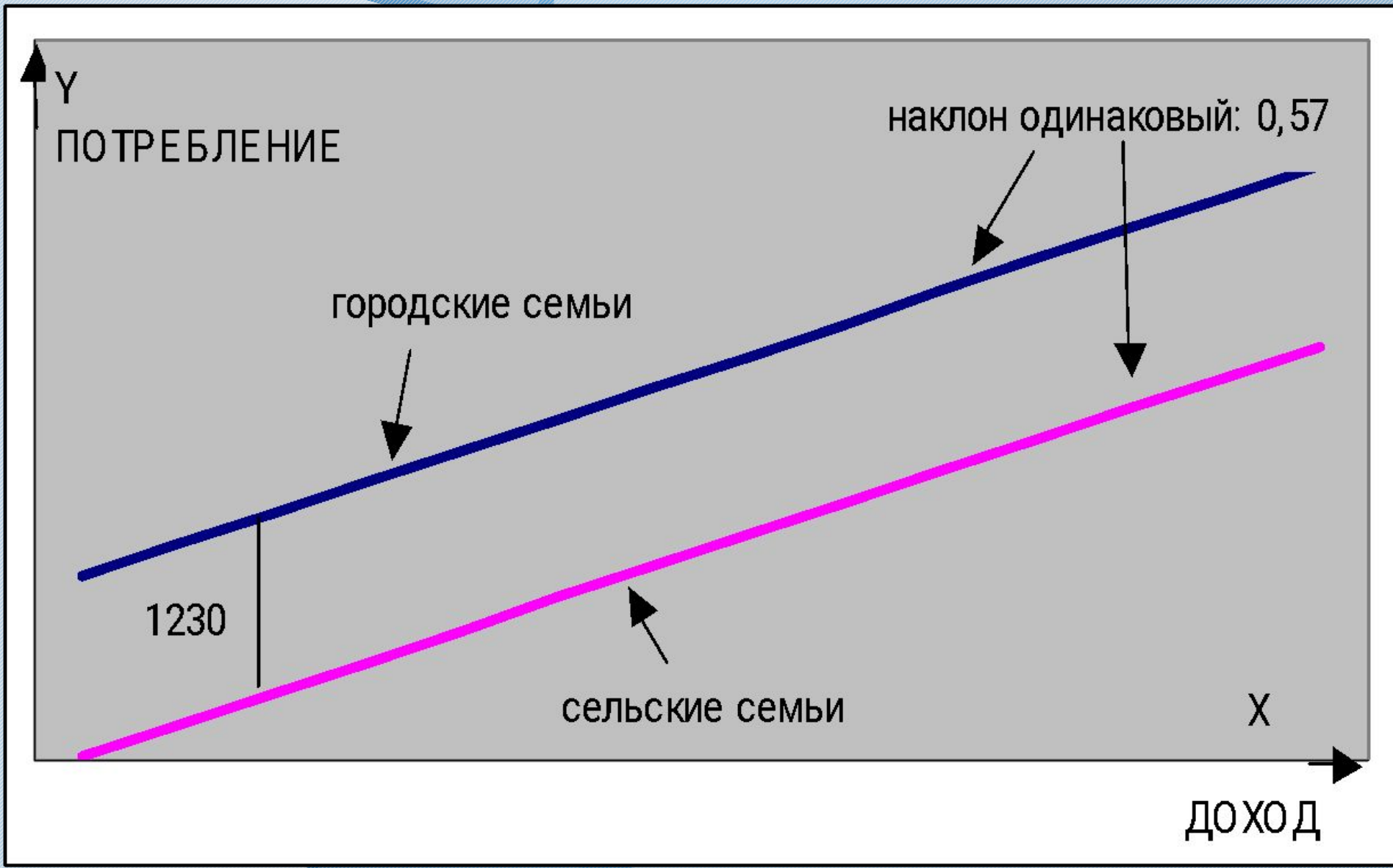
$$\hat{Y} = 3750 + 0{,}57*X$$

Для сельских D=1:

$$\hat{Y} = 3750 + 0{,}57*X - 1230 =$$

$$= 2520 + 0{,}57*X.$$

$$\hat{Y} = 3750 + 0{,}57*X - 1230*D$$



II. ФП для качественного фактора, принимающего более 2-х значений.

Модель без взаимодействия.

Качественный фактор принимает **p** значений (имеет **p** категорий), и

p > 2.

Можно было бы ввести одну ФП, принимающую p различных значений.

Но в этом случае трудно интерпретировать коэффициенты при ФП.

Вводят p ФП, $D_1, D_2, \ldots, D_p$, каждая из которых принимает два значения:

0 и 1.

Каждая такая ФП является индикатором объектов, на которых качественный фактор принимает одно из своих значений.

Одна из ФП объявляется эталонной и в модель не включается.

Т. е. в модель включаются не все p, а только $p-1$ фиктивных переменных.

Эталонной делают ФП – индикатор такой категории (значения качественного признака), с которой хотят сравнивать все остальные $p-1$ категории.

Если, например, эталонной выбрали ФП D_1, то модель имеет вид:

$$Y = \beta_1 + \beta_2 * X_2 + \ldots + \beta_k * X_k + \delta_2 * D_2 + \ldots + \delta_p * D_p + u$$

Если в модель включить все p ФП $D_1, D_2, \ldots, D_p$, то для любого объекта выборки будет выполняться:

$$D_1 + D_2 + \ldots + D_p = 1$$

и будет иметь место совершенная МК $D_1, D_2, \ldots, D_p$ и свободного члена модели.

III. ФП для нескольких качественных факторов.

Модель без взаимодействия.

На **Y** влияют несколько качественных факторов.

Тогда в модель вводят соответствующее количество фиктивных переменных.

ПРИМЕР 5.

Y – з/п работника

X – стаж работника

З\п зависит также от уровня образования сотрудника (4 категории, как и выше) и от его пола.

Для уровня образования, как и выше, вводят 4-е ФП D_1, D_2, D_3, D_4.

Пусть, например, эталонной будет D_3.

Для фактора «пол» вводим ФП П. Пусть, например,

П=0 для мужчин

П=1 для женщин

Модель:

$$Y = \beta_1 + \beta_2 * X + \delta_1 * D_1 + \delta_2 * D_2 + \delta_4 * D_4 + \pi * \Pi + u.$$

IV. Модель со взаимодействием. ФП для коэффициентов наклона.

Для простоты будем рассматривать качественный фактор с 2-я категориями (значениями).

В модели без взаимодействия

$$Y = \beta_1 + \beta_2 * X + \delta * D + u$$

ФП **D** влияет только на значение свободного члена и НЕ влияет на значение коэффициента наклона при X.

Т. е. считается, что качественный фактор:

(а) влияет на значение **Y** для разных категорий объектов, у которых **X** один и тот же;

(б) при изменении фактора X фактор **Y** изменяется ОДИНАКОВО для обеих категорий объектов.

В модели со взаимодействием предположение (б) снимается.

Допускается, что **Y** может по-разному реагировать на изменения X для разных категорий объектов.

Модель со взаимодействием:

$$Y = \beta_1 + \beta_2 * X + \delta * D + \gamma * D * X + u.$$

Ее можно переписать так:

$$Y = (\beta_1 + \delta * D) + (\beta_2 + \gamma * D) * X + u.$$

V. Модель со взаимодействием. Взаимодействие между ФП

ПРИМЕР 8.

Y – з/п сотрудника в рублях,

X – стаж сотрудника, в годах.

На з/п влияют также качественные факторы:

- пол,
- наличие высшего образования.

Вводим ФП П – «пол»:

П = 0 для женщин,

П = 1 для мужчин.

Вводим ФП Е – «наличие высшего образования»:

Е = 0, если в/о нет,

Е = 1, если в/о есть.

Модель:

$$Y = \alpha + \beta * X + \delta * П + \gamma * E + \lambda * П * E + u.$$

Перепишем эту модель в виде:

$$Y = \alpha + \beta * X + (\delta + \lambda * E) * П + \gamma * E + u.$$

Эта модель предполагает, что при постоянном стаже (X) влияние на з/п признака пол (П) различное для групп сотрудников, имеющих и не имеющих высшего образования.

$$Y = \alpha + \beta * X + (\delta + \lambda * E) * П + \gamma * E + u.$$

Т. е. при одинаковом стаже разница в з/п у мужчин (П=1), имеющих в/о (Е=1) и не имеющих в/о (Е=0) составляет ($\gamma + \lambda$) рублей.

При одинаковом стаже разница в з/п у женщин (П=0), имеющих (Е=1) и не имеющих в/о (Е=0) составляет γ рублей.

Модель:

$$Y = \alpha + \beta * X + \delta * П + \gamma * E + \lambda * П * E + u.$$

Эту модель можно переписать по-другому:

$$Y = \alpha + \beta * X + \delta * П + (\gamma + \lambda * П) * E + u.$$

Эта модель предполагает, что при постоянном стаже (X) влияние на з/п наличия или отсутствия в/о различно для мужчин и женщин.

$$Y = \alpha + \beta * X + \delta * П + (\gamma + \lambda * П) * E + u.$$

Т.е. при одинаковом стаже (X) разница в з/п у мужчин (П=1) и женщин (П=0) с в/о (Е=1) составляет $(\delta + \lambda)$ рублей.

При одинаковом стаже (X) разница в з/п у мужчин (П=1) и женщин (П=0) без в/о (Е=0) составляет δ рублей.

$$Y = \alpha + \beta * X + \delta * П + \gamma * E + \lambda * П * E + u.$$

Примечание. Значимость коэффициента **λ** безотносительно к значимости или незначимости остальных коэффициентов при ФП, означает, что имеется значимое различие в з/п категории П = 1, Е = 1 (у нас это мужчины с в/о) над з/п других трех категорий сотрудников при одинаковом стаже.

Критерий Чоу

В практике нередки случаи, когда имеются две выборки пар значений зависимой и объясняющих переменных (X_i; Y_i).

Например, одна выборка пар значений переменных объемом n_1 получена при одних условиях, а другая, объемом n_2 — при несколько измененных условиях. Необходимо выяснить, действительно ли две выборки однородны в регрессионном смысле. Другими словами, можно ли ***объединить*** две выборки в одну и рассматривать единую модель регрессии Y по X?

При достаточных объемах выборок можно было, например, построить интервальные оценки параметров регрессии по каждой из выборок и в случае пересечения соответствующих доверительных интервалов сделать вывод о единой модели регрессии. Возможны и другие подходы.

В случае, если объем хотя бы одной из выборок незначителен, то возможности такого (и аналогичных) подходов резко сужаются из-за невозможности построения сколько-нибудь надежных оценок.

В критерии (тесте) Г. Чоу эти трудности в существенной степени преодолеваются.

Алгоритм теста Чоу:

1.По каждой выборке строятся две линейные регрессионные модели:

$$Y_i = \beta'_0 + \sum_{j=1}^{m} \beta'_j \cdot X_{ij} + \varepsilon'_i \qquad i = \overline{1, n_1}$$

$$Y_i = \beta''_0 + \sum_{j=1}^{m} \beta''_j \cdot X_{ij} + \varepsilon'' \qquad i = \overline{n_1 + 1, n_1 + n_2},$$

Проверяемая нулевая гипотеза имеет вид —

$$H_0: \quad \beta' = \beta''; \; D(\varepsilon') = D(\varepsilon'') = \sigma^2$$

где $\beta' = \beta''$ - векторы параметров двух моделей; $(\varepsilon', \varepsilon'')$ - их случайные возмущения.

2. Рассчитываем суммы квадратов остатков для регрессий по этим подвыборкам

3. Строим регрессию $Y_i = \beta_0 + \sum_{j=1}^{m} \beta_j \cdot X_{ij} + \varepsilon$ по объединенной выборке и рассчитываем ее сумму квадратов остатков

4. Рассчитывается F статистика по формуле:

$$F = \frac{(\sum_{j=1}^{n} e_i^2 - \sum_{j=1}^{n_1} e_i^2 - \sum_{j=1}^{n_2} e_i^2) \cdot (n - 2m - 2)}{(\sum_{j=1}^{n_1} e_i^2 - \sum_{j=1}^{n_2} e_i^2) \cdot (m + 1)}$$

Если $F > F(\alpha, (n-2m-2), (m+1))$, то нулевая гипотеза отвергается и мы не можем объединить две выборки в одну

Если нулевая гипотеза верна, то две регрессионные модели можно объединить в одну объема $n = n_1 + n_2$:

Идея теста Чоу тесно связана с методикой регрессионного анализа с ФП, когда имеется возможность разделения совокупности наблюдений по степени воздействия этого фактора на отдельные группы и требуется установить возможность использования единой модели регрессии.

Оценивание регрессии с использованием ФП более информативно в том отношении, что позволяет использовать t-критерий для оценки существенности влияния каждой фиктивной переменной на зависимую переменную.

Тест Чоу может применяться, например, для выявления стабильности временного ряда. Для этого временной ряд разбивается на две подвыборки: до существенных изменений ряда и после этого. Выдвигается гипотеза о структурной стабильности тенденции ряда и проверяется на основании теста Чоу.