Фиктивные переменные в регрессионных моделях

- В линейную модель множественной регрессии, как правило, включаются количественные факторы *X1, X2,..., Xk*.

- Часто случается так, что отдельные факторы, которые вы хотели бы ввести в регрессионную модель, являются качественными по своей природе и не измеряются числами.

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

Особенности включения в модели регрессии фиктивных переменных.

- Фиктивная переменная — это индикаторная переменная, отражающая качественную характеристику.

- Как правило применяют бинарные фиктивные переменные, которые принимают только два возможных значения: 0 или 1

- При этом 0 означает отсутствие признака у данного объекта; 1- наличие признака.

Пример 3.4.9 Фиктивные переменные сдвига

Орлова И.В., Половников В.А. Экономико-математические методы и модели: компьютерное моделирование: Учеб. пособие – М.: Вузовский учебник

- Построена регрессионная модель зависимости заработной платы работника (Y) от возраста (X) с использованием фиктивной переменной по фактору пол по 20 работникам одного предприятия

$$y = 60,71 + 6,98x + 17,27z$$

- Из полученного уравнения регрессии следует, что при одном и том же возрасте заработная плата у работников мужчин на 17,27$ в месяц выше, чем у женщин.
- Из модели, включающей фиктивную переменную можно получить частные уравнения регрессии для работников мужчин (z=1) и женщин (z=0):

$$y = 77,98 + 6,98x \quad (z = 1)$$

$$y = 60,71 + 6,98x \quad (z = 0).$$

Фиктивные переменные сдвига. Пример 3.4.9.

Орлова И.В., Половников В.А. Экономико-математические методы и модели: компьютерное моделирование: Учеб. пособие – М.: Вузовский учебник

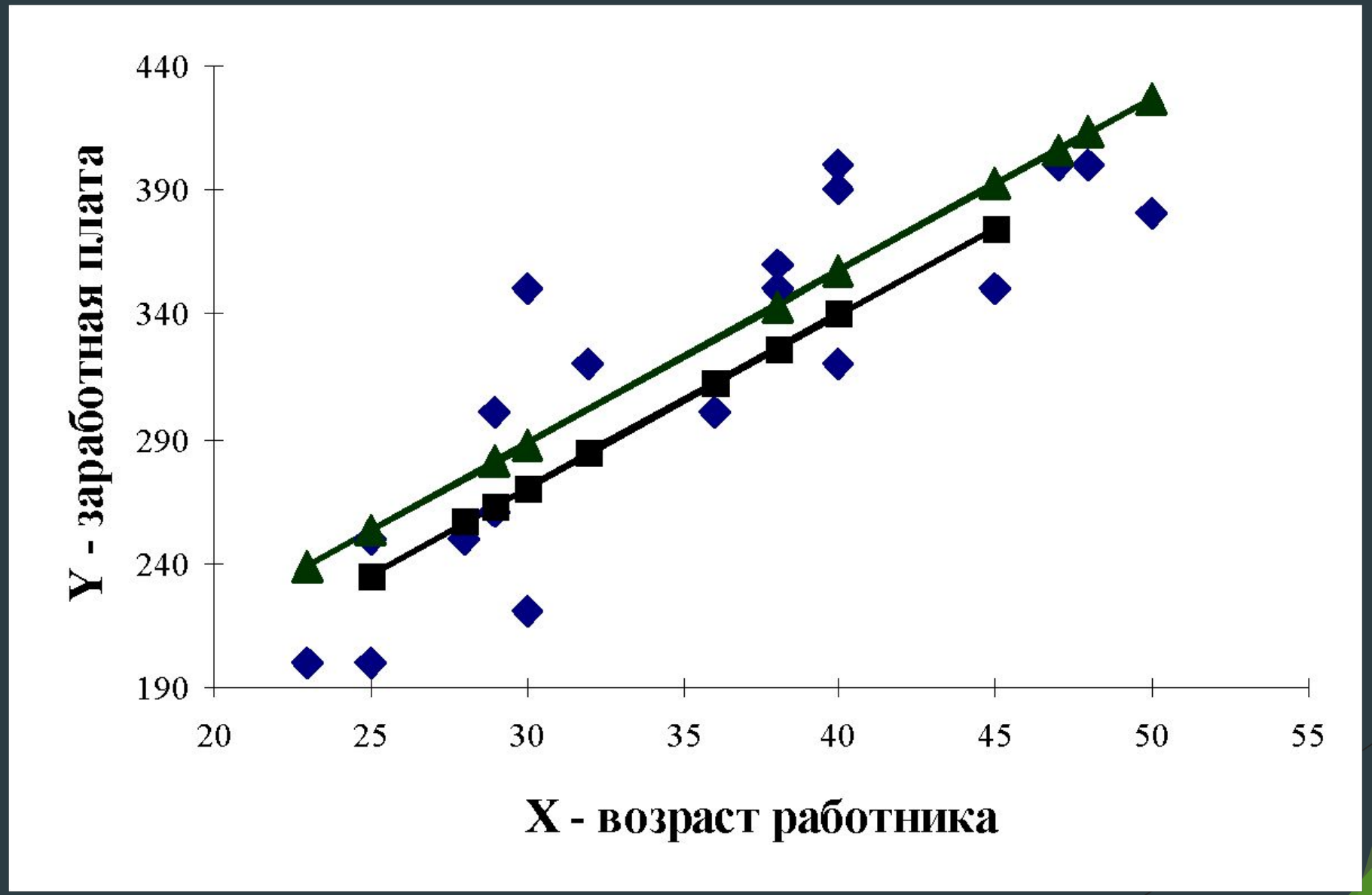
	Коэффициенты	Стандартная ошибка	t-статистика	P-Значение	Нижние 95%	Верхние 95%
Y-пересечение	**60,708**	38,13432111	1,592	0,130	-19,748	141,165
X - возраст (лет)	**6,983**	1,072470179	6,511	0,000	4,720	9,245
Z – пол (1-М, 0-Ж),	**17,275**	17,46232369	**0,989**	**0,336**	**-19,568**	**54,117**

Получили модель

Y=60.708+6.983X+17.275Z

Y=77.983+6.983X - мужчины

Y=60.708+6.983X - женщины


440
390
340
290
240
190
Y - заработная плата
20
25
30
35
40
45
50
55
X - возраст работника

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 о заработной плате мужчин и женщин коэффициент при фиктивной переменной незначим, следовательно, **разницу в оплате труда мужчин и женщин одного возраста можно считать не существенной**.

$$y = 60,71 + 6,98x + 17,27z$$

	Коэффициенты	Стандартная ошибка	t-статистика	P-Значение	Нижние 95%	Верхние 95%
Y-пересечение	**60,708**	38,13432111	1,592	0,130	-19,748	141,165
X - возраст (лет)	**6,983**	1,072470179	6,511	0,000	4,720	9,245
Z - пол (1-М, 0-Ж),	**17,275**	17,46232369	**0,989**	**0,336**	**-19,568**	**54,117**

Использование фиктивных переменных в моделях с временными рядами

- 1) Переменные-индикаторы принадлежности наблюдения к определенному периоду — для моделирования скачкообразных структурных сдвигов. Постоянный структурный сдвиг моделируется переменной равной 0 до определенного момента времени и 1 для всех наблюдений после этого момента времени.
- 2) Сезонные переменные — для моделирования сезонности. Сезонные переменные принимают разные значения в зависимости от того, какому месяцу или кварталу года или какому дню недели соответствует наблюдение.
- 3) Линейный временной тренд — для моделирования постепенных плавных структурных сдвигов. Эта фиктивная переменная показывает, какой промежуток времени прошел от некоторого “нулевого” момента времени до того момента, к которому относится данное наблюдение (координаты данного наблюдения на временной шкале). Если промежутки времени между последовательными наблюдениями одинаковы, то временной тренд можно составить из номеров наблюдений.

Временной тренд отличается от бинарных фиктивных переменных тем, что имеет смысл использовать его степени: t^2 , t^3 и т. д. Они помогают моделировать гладкий, но нелинейный тренд. (Бинарную переменную нет смысла возводить в степень, потому что в результате получится та же самая переменная.)

Например, модель потребления, учитывающая сезонные колебания.

$$y = b_0 + b1x1 + b_2x_2 + b_3x_3,$$

$$x_1 = \begin{cases} 1 & \text{для зимних месяцев} \\ 0 & \text{иначе} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{для весенних месяцев} \\ 0 & \text{иначе} \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{для летних месяцев} \\ 0 & \text{иначе} \end{cases}$$

Следует отметить, что вводить четвертую переменную x_4 для осенних месяцев не требуется, т.к. в этом случае все переменные оказались бы связанными тождеством

$$Xi + X_2 + X_3 + X_4 = 1,$$

что привело бы их к полной коллинеарности и вырожденности информационной матрицы $(X^T X)$.

Модель с фиктивными переменными имеет вид:

- $y = f(x_1, \ldots, x_p, z_{11}, z_{12}, \ldots, z_{21}, z_{22}, \ldots, z_{j1}, z_{j2}, \ldots, \varepsilon)$,

- где y – зависимая переменная; $x_1, \ldots, x_p$ – количественные независимые переменные; z_{11}, z_{12} – фиктивные переменные, соответствующие категориям первого неколичественного показателя; z_{21}, z_{22} – фиктивные переменные, соответствующие категориям второго неколичественного показателя; z_{j1}, z_{j2} – фиктивные переменные, соответствующие категориям j-ого неколичественного показателя; ε – случайный остаток.

Фиктивные переменные наклона.

- Возможна комбинация фиктивных переменных различных видов. Она позволяет моделировать ***изменение наклона тренда*** с определенного момента. Помимо тренда, в регрессию тогда вводится следующая переменная: в начале выборки до некоторого момента времени она равна 0, а далее она представляет собой временной тренд.

- С помощью фиктивных переменных можно строить и оценивать кусочно-линейные модели, которые применяются для исследования структурных изменений.

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

Результаты регрессионного анализа

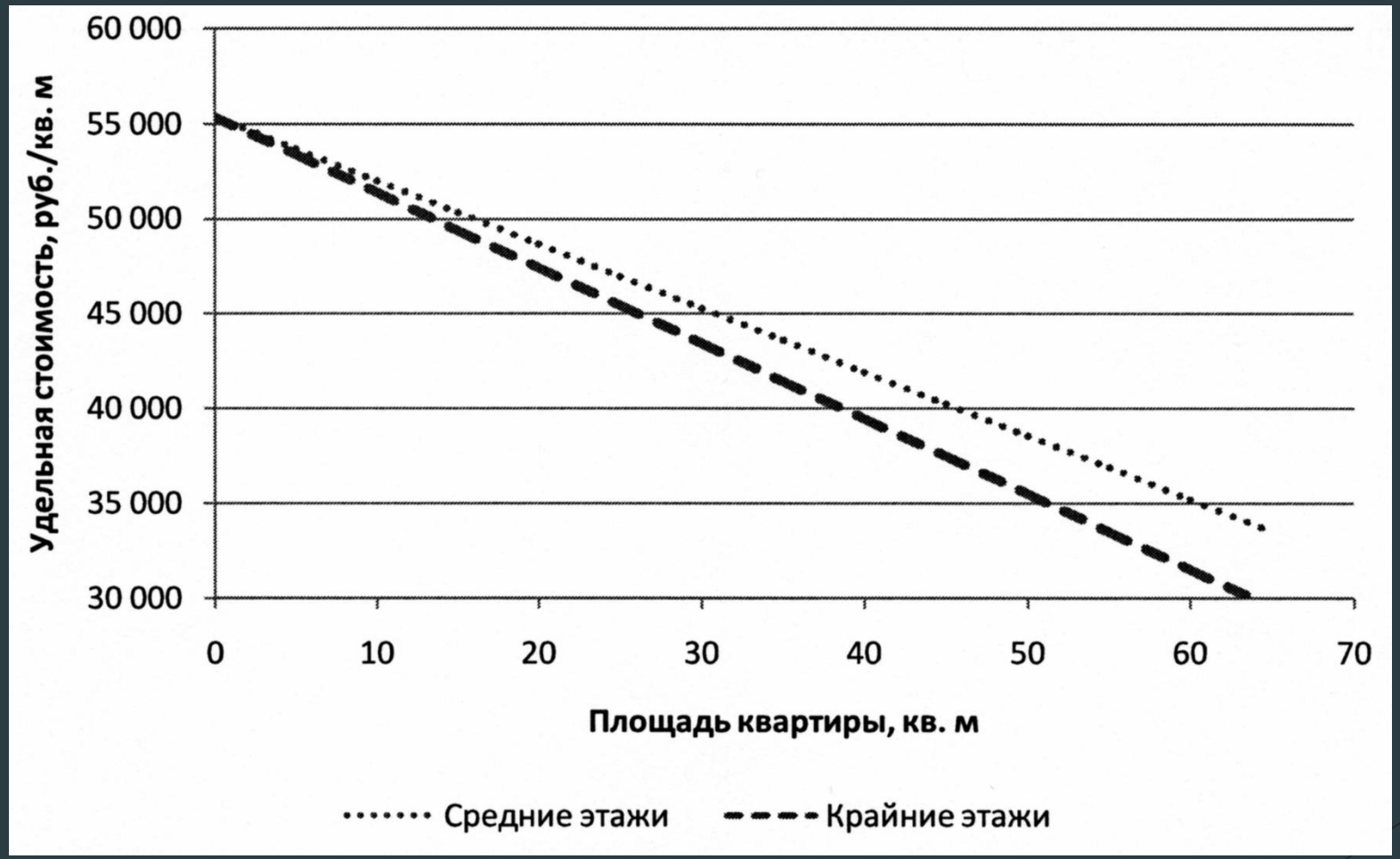
Регрессионная статистика	
Множественный R	0,829
R-квадрат	0,687
Нормированный R-квадрат	0,683
Стандартная ошибка	2 573
Наблюдения	146

Дисперсионный анализ

	df	SS	*MS*	*F*	*Значимость F*
Регрессия	2	2 082 406 181	1 041 203 091	157	0,0000
Остаток	143	946 894 802	6 621 642		
Итого	145	3 029 300 983			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	53 358	1 016	53	0,00000	51 787	55 803
Средний этаж	61	9	7	0,00000	43	79
Общая площадь, кв. м	-398	22	-18	0,00000	-442	-353

Визуализация построенной регрессионной модели с использованием переменной наклона.



Фактически полученная модель:

$$Y = a1 * X1 * S + a2 * S + c$$

идентична двум моделям:

$Y = (a1 + a2) * S + c = -336 * S + 55\,358$ для квартир на средних этажах

$Y = (a2) * S + c = -398 * S + 55\,358$ для квартир на крайних этажах

В данном случае в зависимости от значения качественной переменной изменяется коэффициент при количественном параметре, т.е. меняется наклон графика линии регрессии.