А.А. Марчук

Русская Служба Оценки (www.rusvs.ru)

Учет качественных параметров в регрессионном анализе

Применение статистики в оценке имущества является одной из наиболее «модных» тем в оценочном сообществе. Статистические процедуры, давно известные математикам и специалистам по эконометрике, могут быть полезны оценщику, позволяя сделать расчеты более надежными и обоснованными. Развитие информационной открытости рынков, унификация баз данных, давно обсуждаемый доступ оценщиков к данным регистрирующих органов и т.п., безусловно, потребуют от оценщиков применение самых современных способов обработки и интерпретации информации.

Одним из наиболее эффективных и полезных для оценщика статистических инструментов является регрессионный анализ. Регрессионный анализ позволяет выявить ценообразующие факторы и определить вид зависимости. Но при применении данного инструмента оценщики сталкиваются с целым рядом проблем (недостаточное число аналогов, недоступность информации по объектам-аналогам, волатильность рынков и другие). Эта статья посвящена проблеме учета качественных параметров при построении регрессионных зависимостей.

Параметры, описывающие характеристики того или иного объекта оценки, можно разделить на два типа: количественные и качественные. Количественные параметры – параметры, значения которых выражаются числом. Примерами таких параметров могут выступать, например, площадь объекта, мощность двигателя, выручка компании и т.п.

Но описать полностью тот или иной объект только количественными параметрами удается далеко не всегда. Существенная часть характеристик оцениваемых объектов носит качественный характер, т.е. описываются качественными параметрами. Значение качественной переменной выражается текстовым описанием, рисунком или каким-либо другим поясняющим его смысл способом. Измерить качественные параметры при помощи числовой шкалы невозможно. Примерами таких переменных могут выступать класс объекта, материал изготовления, тип привода, район расположения объекта и т.п.

Учет качественных параметров при регрессионном анализе возможен различными способами. Наиболее простое решение – построение индивидуальных моделей для каждого значения (градации) качественного параметра. На первый взгляд такое решение является крайне эффективным – чем меньше параметров учитывается в модели, тем проще решение задачи. Но, на самом деле, такое решение далеко не всегда самое лучшее. Представьте, что для оценки объекта необходимо будет подобрать аналоги, удовлетворяющие следующим критериям: однокомнатные квартиры, расположенные на первом этаже в домах типа «хрущевка» в определенном районе города, внутри жилого квартала и с отделкой определенного уровня. Даже на развитом рынке (например, рынок жилья городов-миллионников) поиск аналогов, удовлетворяющих столь жестким требованиям, скорей всего, не увенчается успехом. Кроме того, с уменьшением количества аналогов надежность оценок коэффициентов модели существенно снижается, а погрешность полученной модели, соответственно, возрастает.

Альтернативным вариантом является учет качественных параметров в модели, для чего используются фиктивные переменные (в литературе также встречается термин «переменные – манекены»). Данный способ является более эффективным, т.к. появляется возможность оценить статистическую значимость влияния данного фактора на зависимую переменную на фоне других параметров, включенных в модель, и повысить надежность модели за счет включения большего количества аналогов.

Фиктивная переменная (dummy variable) — в эконометрике переменная модели, полученная путем преобразования (напр., с помощью балльных оценок) информации, содержащей качественные и другие не поддающиеся числовой оценке величины. Ф. п. используются как простое средство для включения подобной информации в регрессионный анализ. Напр., добавление Ф. п., принимающей только два значения — 0 и 1 в качестве дополнительной объясняющей переменной, часто используется при анализе сезонных колебаний. [1]

Фиктивные переменные подразделяются на переменные сдвига и переменные наклона. Выбор в пользу того или иного типа переменных (или их комбинаций) является «содержательной» задачей и зависит от характера используемых в модели параметров. Далее продемонстрированы особенности данных видов переменных на нескольких примерах.

1. Переменные сдвига

1.1. Пример №1: модель с одной фиктивной переменной.

Рассмотрим самый простой случай регрессии – учтем в модели только одну, качественную переменную. Данный пример является упрощенным вариантом применения регрессионного анализа и предназначен для демонстрации сути фиктивных переменных.

В нижеследующей таблице представлены данные по ценам предложения однокомнатных квартир одного типа (панельные 5-ти этажные дома), расположенные в центральном районе г. Архангельска[1].

Табл. 1. Информация об аналогах для Примера №1

№, п./п.	Дата	Кол-во комнат	Район	Этаж	Этаж-ность	Площадь общая, кв. м	Жилая площадь, кв. м	Кухня, кв. м	Тип дома	Удельная цена, руб./кв. м	«Средний этаж»
1	2	3	4	5	6	7	8	9	10	11	12
1	20.06.09	1	Центр	1	5	31,3	18,2	6	пан.	41 534	0
2	16.05.09	1	Центр	1	5	31	18	6	пан.	41 613	0
3	25.09.09	1	Центр	1	5	31	18	6,5	пан.	41 613	0
4	16.10.09	1	Центр	5	5	31	17,5	6	пан.	41 935	0
5	13.11.09	1	Центр	5	5	30,7	17	5,6	пан.	42 345	0
6	15.08.09	1	Центр	1	5	31	17	5	пан.	42 903	0
7	13.11.09	1	Центр	5	5	31	17,5	5,7	пан.	43 548	0
8	23.10.09	1	Центр	1	5	30	17,5	5,5	пан.	45 000	0
9	20.11.09	1	Центр	5	5	30	18	5,6	пан.	45 000	0
10	16.05.09	1	Центр	1	5	29,3	17,5	6,1	пан.	45 000	0
11	20.11.09	1	Центр	1	5	30	16,4	6,3	пан.	45 000	0
12	02.10.09	1	Центр	1	5	30	18	6	пан.	45 000	0
13	01.08.09	1	Центр	5	5	30,2	17,4	5,4	пан.	45 000	0
14	11.09.09	1	Центр	3	5	32	18	6	пан.	43 750	1
15	19.09.09	1	Центр	3	5	30,5	18,6	5,4	пан.	44 262	1
16	16.10.09	1	Центр	3	5	31	18,5	6	пан.	44 839	1
17	28.08.09	1	Центр	4	5	31	18	5	пан.	45 161	1
18	02.10.09	1	Центр	4	5	31	17	6,7	пан.	45 161	1
19	13.11.09	1	Центр	3	5	30,4	16,6	6,3	пан.	46 053	1
20	09.10.09	1	Центр	2	5	32	17	6	пан.	46 875	1
21	30.10.09	1	Центр	2	5	30	17,1	5,4	пан.	48 333	1
22	19.09.09	1	Центр	2	5	30	17,1	5,4	пан.	49 667	1
23	16.10.09	1	Центр	2	5	30	17	5,6	пан.	50 000	1
24	09.10.09	1	Центр	4	5	31	18	6	пан.	50 000	1

Как видно из таблицы, объекты достаточно однородны по площади, расположены в одном районе. При этом аналоги №№1…13 расположены на крайних этажах, 14-24 – на средних этажах. Общеизвестный факт, что квартиры на крайних этажах обычно стоят дешевле аналогичных квартир на средних этажах. Анализ данных, представленных в Табл. 1, это подтверждает – см. Табл. 2.

[1] Данные предоставлены Агентством недвижимости «Троицкий Дом» (г. Архангельск, www.3dom.ru)

Табл. 2. Средние значения

Этаж	Показатель	Удельная цена, руб./ кв. м
Крайние этажи	Минимальное значение	41 534
	Максимальное значение	45 000
	Среднее	**43 499**
Средние этажи	Минимальное значение	43 750
	Максимальное значение	50 000
	Среднее	**46 736**
По всей выборке	Минимальное значение	41 534
	Максимальное значение	50 000
	Среднее	**44 983**

Учтем различие «крайний этаж» / «средний этаж» при построении модели, для чего введем фиктивную переменную X_1 («Средний этаж», см. колонку 12 в Табл. 1), которая будет принимать следующие значения:

$X_1 = 1$ для квартир, расположенных на средних этажах;

$X_1 = 0$ для квартир, расположенных на крайних этажах.

Искомая модель будет иметь вид:

$$\mathbf{Y = a_1 * X_1 + c} \qquad \textbf{(модель №1)}$$

где:

Y - удельная стоимость;

X_1 - фиктивная переменная «Средний этаж»;

a_1 - коэффициент модели;

c - константа.

Полученные результаты представлены в Табл. 3 (строка №1). Сопоставляя Табл. 2 и Табл. 3 легко заметить, что константа модели «c» равна среднему значению удельной стоимости квартир, расположенных на крайних этажах, а коэффициент при фиктивной переменной «Средний этаж» равен разнице между средними удельными стоимостями квартир на средних и крайних этажах:

$$\mathbf{a_1 = 46\,736 - 43\,499 = 3\,237}$$

Фактически коэффициент при фиктивной переменной «Средний этаж» отвечает на вопрос:

«На сколько в среднем квартиры на средних этажах дороже квартир на крайних этажах?».

Аналогичную модель можно построить с фиктивной переменной X_2 («Крайний этаж»):

$$\mathbf{Y = a_2 * X_2 + c} \qquad \textbf{(модель №2)}$$

где:

X_2 - фиктивная переменная «Крайний этаж», принимающая следующие значения:

$X_2 = 0$ для квартир, расположенных на средних этажах;

$X_2 = 1$ для квартир, расположенных на крайних этажах.

Результаты расчетов представлены в Табл. 3 (строка №2). Данная модель идентична модели №1, но константа модели равна среднему значению удельной стоимости квартир на средних этажах, а коэффициент при фиктивной переменной «Крайний этаж» отвечает на вопрос:

«На сколько в среднем квартиры на крайних этажах дешевле квартир на средних этажах?».

На этих же данных возможно построить еще один вариант модели, куда будут включены обе обозначенные выше фиктивные переменные. При этом будет наблюдаться полная мультиколлинеарность[2] (параметры X_2 и X_1 связаны выражением $\mathbf{X_2 = 1- X_1}$), для устранения которой необходимо исключить из спецификации модели константу:

$$Y=a_1*X_1+a_2*X_2$$

Результаты также представлены в Табл. 3 (строка №3). Коэффициенты при фиктивных переменных в данной модели равны средним значениям стоимости квартир на средних и крайних этажах соответственно.

Табл. 3. Результаты регрессионного анализа

Модель	a_1	a_2	C	R^2
Модель №1 $Y=a_1*X_1+c$	3 237 (546)	–	43 499 (806)	0,42
Модель №2 $Y=a_2*X_2+c$	–	-3 237 (546)	46 736 (806)	0,42
Модель №3 $Y=a_1*X_1+a_2*X_2$	46 736 (593)	43 499 (546)	–	0,998

Примечания:

1. *Следует отметить достаточно низкое значение R^2. Это объясняется тем, что анализируемые квартиры отличаются не только этажом расположения, но и состоянием, а также местоположением внутри центрального района города. Задачей этого и последующих примеров является демонстрация сути фиктивных переменных;*
2. *В скобках указаны стандартные ошибки для полученных коэффициентов модели;*
3. *Для модели без константы (модель №3) вместо коэффициента детерминации R^2 определяется нецентрированный R^2. Сопоставление нецентрированного R^2 с коэффициентом детерминации некорректно.*

1.2. Пример №2: модель с одной фиктивной и одной количественной переменными

Пример №1 является достаточно простым случаем. Рассмотрим более сложный вариант.

На графике ниже представлена информация о ценах предложения квартир в г. Архангельске. В данную выборку были включены такие же квартиры, как и в примере №1 (т.е. панельные 5-ти этажные дома, расположенные в одном районе города), но в данном случае отсутствовало ограничение по количеству комнат.

Из графика видно, что в выборку были включены квартиры разной площади: однокомнатные площадью около 30 кв. м, 2-х комнатные площадью 40-48 кв. м и 3-х и 4-х комнатные площадью более 55 кв. м (на графике видно три группы аналогов). Также на графике видно, что квартиры на средних этажах обычно чуть дороже квартир на крайних этажах.

[2] Мультиколлинеарность [multicollinearity] — понятие математической статистики — тесная корреляционная взаимосвязь между отбираемыми для анализа факторами, совместно воздействующими на общий результат. Эта связь затрудняет оценивание параметров регрессии в частности, при анализе эконометрической модели. [1]

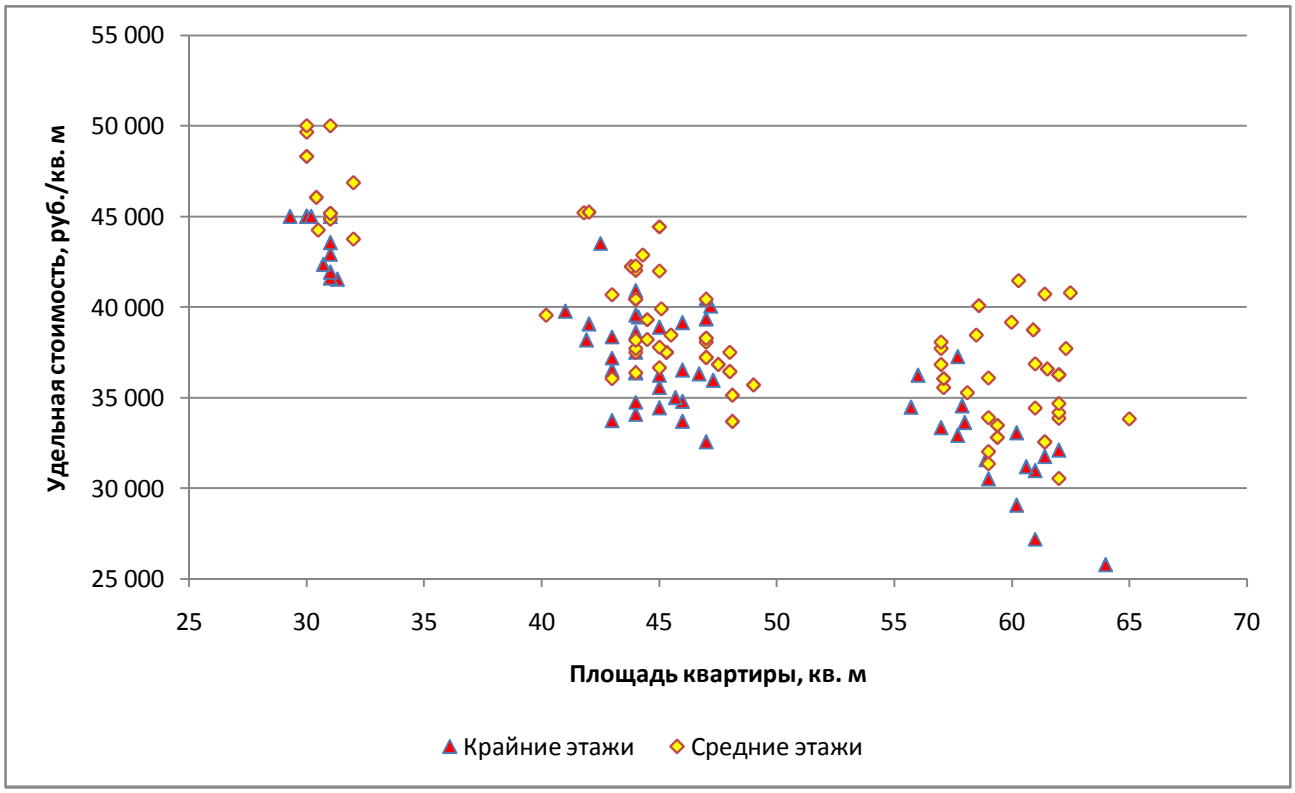

Рис. 1. Информация об аналогах для Примера №2

Попробуем учесть различие в площади объектов и этаже расположения в модели, для чего построим модель вида:

$$Y=a_1*X_1+ a_2*S + c$$

Где:

S - общая площадь квартиры;

X_1 - фиктивная переменная средний этаж.

Результаты анализа представлены в нижеследующей таблице (формат таблицы соответствует результатам действия надстройки «Регрессия» из «Пакета анализа» MS Excel).

Табл. 4. Результаты регрессионного анализа

Регрессионная статистика	
Множественный R	0,825
R-квадрат	0,681
Нормированный R-квадрат	0,676
Стандартная ошибка	2 600
Наблюдения	146

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	2	2 062 718 366	1 031 359 183	153	0,00000
Остаток	143	966 582 616	6 759 319		
Итого	145	3 029 300 983			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	53 795	1 016	53	0,00000	51 787	55 803
Средний этаж	2 859	440	6	0,00000	1 989	3 729
Общая площадь, кв. м	-363	21	-17	0,00000	-405	-321

Графическое представление полученных результатов представлено на Рис. 2. Как видно из рисунка, стоимость квартир на крайних этажах ниже стоимости квартир на средних этажах. Разница в стоимости равна коэффициенту при фиктивной переменной и в данном случае составляет 2 859 руб./ кв. м. При этом эта разница одинакова для квартир разной площади (на графике линии, соответствующие средним и крайним этажам параллельны).

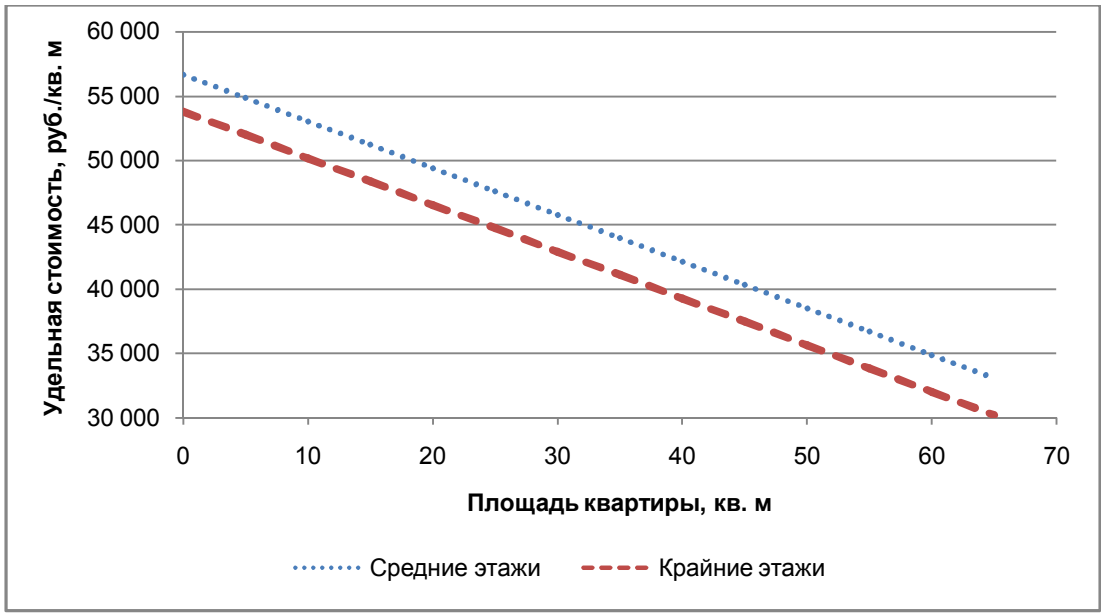

Рис. 2. Визуализация построенной регрессионной модели

Переменные, аналогичные использованным в моделях переменным «Средний этаж» и «Крайний этаж», принято называть переменными сдвига и, использовать, когда исследователь предполагает, что влияние данного качественного параметра на зависимую переменную одинаково для всех значений количественной переменной. Фактически речь идет о параллельном сдвиге базовой модели: при неизменном коэффициенте при количественной переменной меняется константа модели.

2. Переменные наклона

Как было указано выше, переменные сдвига применяются, когда влияние качественного параметра постоянно для любых наблюдений (аналогов). При этом на практике такое встречается далеко не всегда. Если рассмотреть те же квартиры, то можно предположить, что для квартир с разным количеством комнат разница в удельной стоимости будет различной. Для учета таких случаев применяются фиктивные переменные наклона.

Попытаемся улучшить модель, построенную по результатам примера №2, для чего заменим фиктивную переменную «Средний этаж» следующей переменной $\mathbf{Z_1}$:

$$\mathbf{Z_1 = X_1 * S}$$

где:

X_1 - фиктивная переменная «Средний этаж» (см. выше);

S - площадь квартиры.

Модель в этом случае будет выглядеть следующим образом:

$$\mathbf{Y = a_1 * Z_1 + a_2 * S + c = a_1 * X_1 * S + a_2 * S + c}$$

Результаты расчетов коэффициентом и статистик модели представлены в Табл. 5.

Табл. 5. 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	0,00000
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	55 358	1 018	54	0,00000	53 347	57 370
Общая площадь, кв. м	-398	22	-18	0,00000	-442	-353
Средний этаж * Общая площадь	61	9	7	0,00000	43	79

На графике данная зависимость будет выглядеть следующим образом:

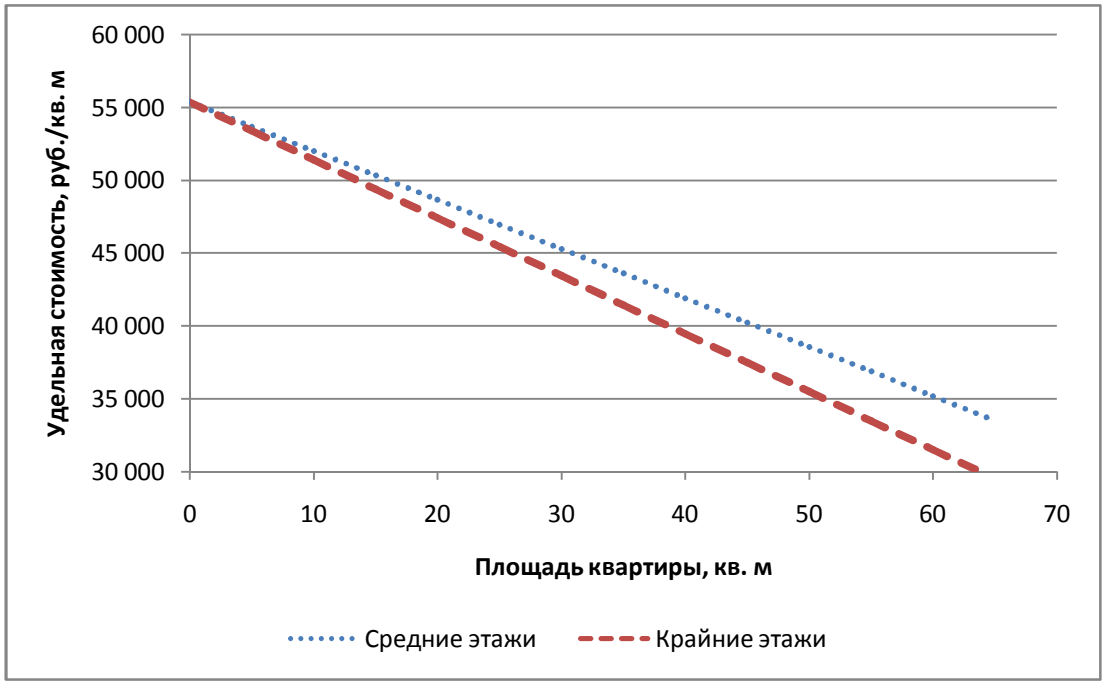

Рис. 3. Визуализация построенной регрессионной модели с использованием переменной наклона

Как видно из Рис. 3, линии, соответствующие средним и крайним этажам, в данном случае не параллельны. Фактически полученная модель:

$$Y = a_1 * X_1 * S + a_2 * S + c$$

идентична двум моделям:

$Y = (a_1 + a_2) * S + c = -336 * S + 55\,358$ для квартир на средних этажах

$Y = (a_2) * S + c = -398 * S + 55\,358$ 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. При этом константа модели остается постоянной для различных значений качественного параметра.

Комбинируя фиктивные переменные наклона и переменные сдвига можно получить модель, полностью идентичную индивидуальным моделям для разных значений количественной переменной.

Например, если по тем же данным построить отдельные модели для квартир на различных этажах, получатся следующие модели:

Для квартир на крайних этажах: **$y = -404,5*S + 55\,688$** ($R^2 = 0,776$)

Для квартир на средних этажах: **$y = -330,4*S + 55\,060$** ($R^2 = 0,592$)

Аналогичные результаты получаются, если построить общую модель для всех этажей расположения вида:

$$Y = a_1*X_1 + a_2*X_1*S + a_3*S + c$$

$$Y = -628 * X_1 + 74,1 * X_1 * S + (-404,5) * S + 55\,688 \ (R^2 = 0,688)$$

Легко заметить, что данная модель при $X_1=1$ превращается в модель для квартир на средних этажах, а при $X_1=0$ – в модель для квартир на крайних этажах, указанных выше.

3. Учет нескольких градаций значений качественного параметра

В представленных выше примерах рассматривались варианты, когда качественный параметр принимает только два значения: «Крайние этажи» или «Средние этажи». При этом качественные параметры, естественно, бывают и более «сложными» и могут принимать различные значения. Например, такой качественный параметр, как «Район города» может принимать большое количество значений.

Учет таких качественных параметров при построении регрессионных моделей может быть выполнен различными способами. Одним из вариантов является замена данного качественного параметра комплексом бинарных переменных: качественный параметр, принимающий n вариантов значений (градаций), заменяется (n-1) бинарным признаком.

Например, если в выборке представлены квартиры 5-ти районов города, то комплекс бинарных признаков будет выглядеть следующим образом:

Табл. 6. Замена качественных параметров бинарными признаками

Район города	Бинарные признаки			
	Район №1	Район №2	Район №3	Район №4
Район №1	1	0	0	0
Район №2	0	1	0	0
Район №3	0	0	1	0
Район №4	0	0	0	1
Район №5	0	0	0	0

Возвратимся к примерам, описанным выше. Анализировалась разница в стоимости квартир, расположенных на крайних и средних этажах. Можно предположить, что такая разница будет неодинаковой для квартир с различным количеством комнат. В выборке представлены квартиры 1-2-3-4-х комнатные. Попробуем учесть различие в стоимости квартир для различного количества комнат путем добавления качественной переменной «Этаж расположения» со следующими градациями:

- Однокомнатные квартиры, расположенные на средних этажах;
- Двухкомнатные квартиры, расположенные на средних этажах;

- Трехкомнатные квартиры, расположенные на средних этажах;
- Четырехкомнатные квартиры, расположенные на средних этажах;
- Квартиры, расположенные на крайних этажах.

Табл. 7. Учет бинарными признаками этажа расположения квартиры

Этаж расположения	Бинарные признаки			
	X_1	X_2	X_3	X_4
Однокомнатные квартиры, средние этажи	1	0	0	0
Двухкомнатные квартиры, средние этажи	0	1	0	0
Трехкомнатные квартиры, средние этажи	0	0	1	0
Четырехкомнатные квартиры, средние этажи	0	0	0	1
Квартиры, расположенные на крайних этажах	0	0	0	0

Модель в данном случае будет выглядеть следующим образом:

$$Y = a_1*X_1 + a_2*X_2 + a_3*X_3 + a_4*X_4 + a_5*S + c$$

Результаты регрессионного анализа представлены в Табл. 8.

Табл. 8. Результаты регрессионного анализа

Регрессионная статистика	
Множественный R	0,842
R-квадрат	0,709
Нормированный R-квадрат	0,699
Стандартная ошибка	2 509
Наблюдения	146

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	5	2 148 083 689	429 616 738	68	0,00000
Остаток	140	881 217 293	6 294 409		
Итого	145	3 029 300 983			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
C	53 705	1 320	41	0,00000	51 095	56 314
a_5	-361	28	-13	0,00000	-417	-305
a_1	4 150	916	5	0,00001	2 339	5 962
a_2	2 117	575	4	0,00033	979	3 254
a_3	2 499	602	4	0,00006	1 309	3 689
a_4	5 465	1 095	5	0,00000	3 301	7 629

Такой подход позволяет не только выявить различия во влиянии отдельных градаций качественного параметра, но и сделать вывод о статистической значимости различия между разными градациями признака. Например, из Табл. 8 видно, что коэффициенты при 2-х и 3-х комнатных квартирах близки. Проверим гипотезу о том, что $a_2 = a_3$.

Для проверки данной гипотезы рассчитаем статистику Стьюдента, полученную величину сравним с критическим значением[3]:

[3] Для нахождения критического значения статистики Стьюдента необходимо воспользоваться статистическими таблицами или использовать функцию MS Excel СТЬЮДРАСПОБР()

$$t = \frac{|2\,117 - 2\,499|}{\max(602; 575)} = 0,635 < t_{кр} = 1,997$$

Т.к. $t<t_{кр}$, делаем вывод о том, что разница в удельной стоимости квартир на средних этажах между 2-х и 3-х комнатными квартирами статистически не значима (при прочих равных условиях). Поэтому без ущерба для точности модели переменные X_2 и X_3 можно объединить в одну[4].

4. Заключение

Подавляющее большинство объектов оценки нельзя описать только при помощи количественных параметров. Такие параметры, как класс объекта или его состояние, местоположение объекта, материал основных элементов и многие другие зачастую оказывают существенное влияние на стоимость. Такие параметры принято называть качественными. При проведении регрессионного анализа необходимо учитывать различия в качественных параметрах, для чего в модель обычно включают одну или несколько фиктивных переменных.

Фиктивные переменные подразделяются на переменные сдвига и переменные наклона. Переменные сдвига позволяют учесть влияние качественного параметра в случае, если данный параметр оказывает «одинаковое» влияние на зависимую переменную (в большинстве оценочных задач в качестве зависимой переменной выступает стоимость или удельная стоимость).

Если оценщик подозревает, что с ростом того или иного количественного параметра влияние качественной переменной изменяется, следует отдать предпочтение фиктивным переменным другого типа – переменным наклона.

5. Литература

1. Лопатников Л. И.Экономико-математический словарь: Словарь современной экономической науки. — 5-е изд., перераб. и доп. — М.: Дело, 2003. — 520 с.
2. Ю.Н. Тюрин, А.А. Макаров Анализ данных на компьютере / Под. ред. В.Э.Фигурнова. - 3-е изд., перераб. и доп. – М.:ИНФРА-М, 2003
3. Вербик Марно. Путеводитель по современной эконометрике. Пер. с англ. В.А. Банникова. Научн. ред. и предисл. С.А. Айвазяна. – М.: Научная книга, 2008. – 616 с.

[4] Т.е. заменить переменные X_2 и X_3 одной переменной $X_{2,3} = X_2 + X_3$