

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ
ФЕДЕРАЦИИ

ФГБОУ ВО «Уральский государственный экономический университет»

В.П. Часовских

Прикладные эконометрические модели на макро- и микроуровне

38.06.01 - Экономика профиль Экономика и управление народным хозяйством
Линейная регрессия. Ранговая корреляция. Множественная регрессия.

Екатеринбург 2021

Линейная регрессия

Линейная регрессия является самым простым подходом для статистического обучения и анализа. Понимание этой простой модели создаст хорошую базу, прежде чем перейти к более сложным подходам.

Линейная регрессия позволяет ответить на следующие вопросы:

- Есть ли связь между 2 переменными?
- Насколько прочны отношения?
- Какая переменная вносит наибольший вклад?
- Насколько точно мы можем оценить влияние каждой переменной?
- Насколько точно мы можем предсказать цель?
- Являются ли отношения линейными?
- Есть ли эффект взаимодействия?

Предположим, у нас есть только одна переменная x и одна цель – y . Тогда линейная регрессия выражается как:

$$y = a_0 + a_1x_1$$

Уравнение для линейной модели с 1 переменной и 1 целью.

В приведенном выше уравнении a_0 и a_1 являются коэффициентами. Эти коэффициенты – то, что нам нужно, чтобы делать прогнозы с нашей моделью.

Итак, как мы можем найти эти параметры?

Чтобы найти параметры, нам нужно минимизировать **сумму квадратов ошибок**. Конечно, линейная модель не идеальна, и она не будет точно предсказывать все данные, а это означает, что существует разница между фактическим значением и прогнозом. Ошибка легко вычисляется с помощью:

$$e_i = y_i - \hat{y}_i$$

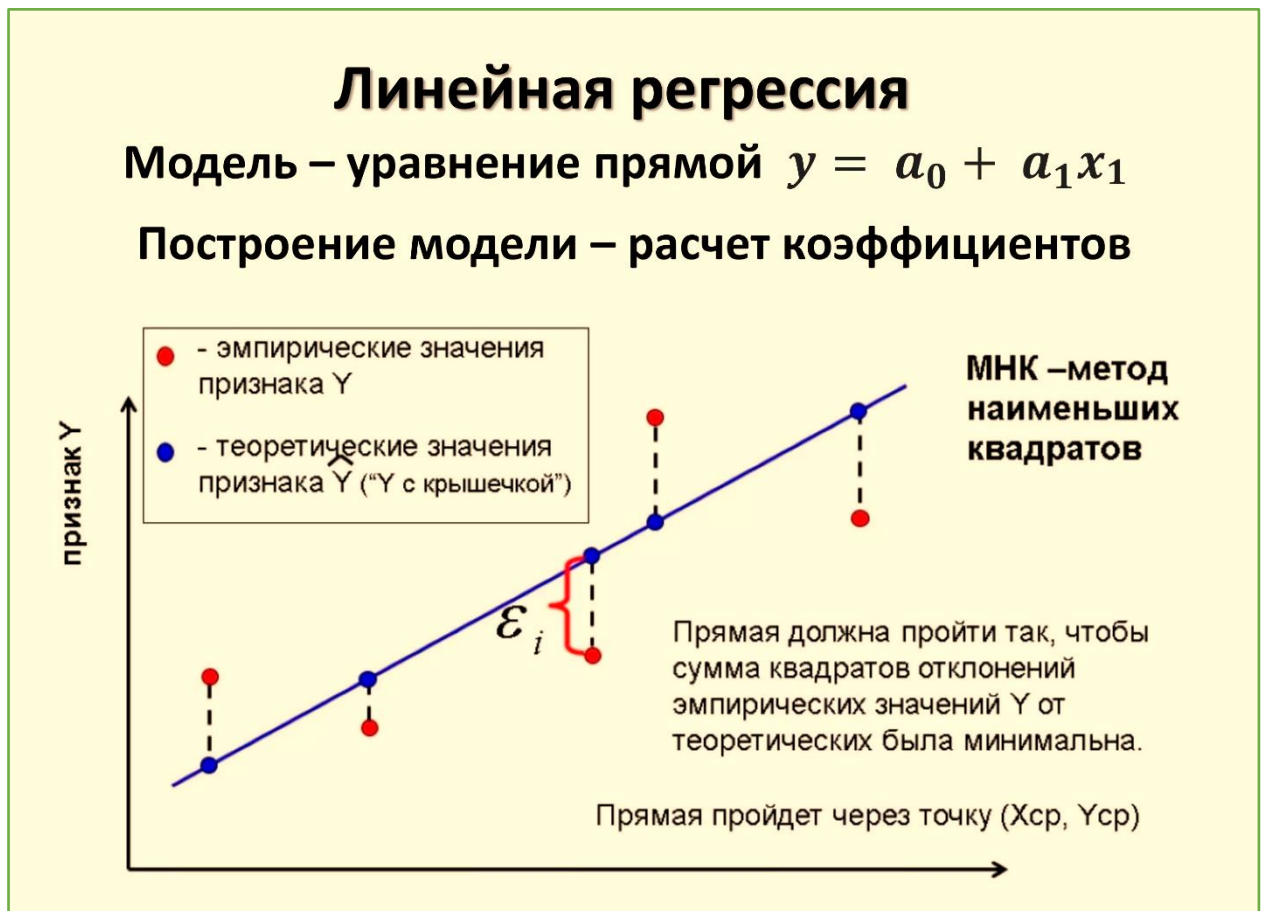
т.е. из истинного значения y_i вычитается прогноз \hat{y}_i

Но почему ошибки возводятся в квадрат?

Принято ошибку возводить в квадрат, потому что прогноз может быть выше или ниже истинного значения, что приводит к отрицательной или положительной разнице соответственно. Если бы мы не возводили в квадрат ошибки, сумма ошибок могла бы уменьшиться из-за отрицательных различий, а не потому, что модель хорошо подходит.

Кроме того, возведение в квадрат ошибок учитывает большие различия, поэтому минимизация квадратов ошибок «гарантирует» лучшую модель.

Рассмотрим график линейной регрессии чтобы лучше понять.



На приведенном выше графике красные точки — это истинные данные, а синяя линия - линейная модель. Прерывистые черные линии иллюстрируют ошибки между предсказанными и истинными значениями. Таким образом, синяя линия - это та, которая минимизирует сумму квадратов длины прерывистых черных линий.

Математический вывод коэффициентов определяет следующие выражения:

$$a_i = \frac{\sum_{i=1}^n (x_i - \hat{x})(y_i - \hat{y})}{\sum_{i=1}^n (x_i - \hat{x})^2}$$

$$a_0 = \hat{y} - a_1 \hat{x}$$

Где \hat{x} и \hat{y} представляют собой среднее.

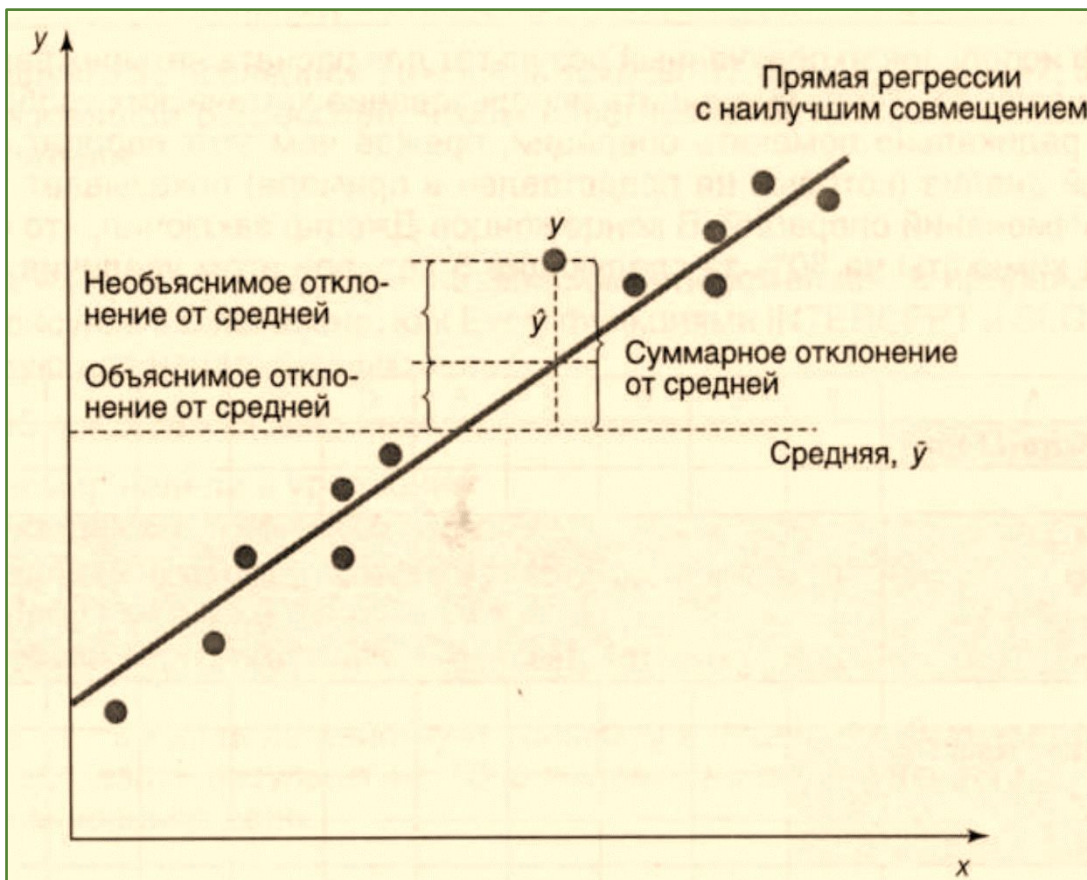
Как определить, что коэффициенты актуальны для нашего прогнозирования?
Для этого необходимо оценить корреляцию.

Коэффициент детерминации

Предположим, что есть ряд наблюдений y_i , рассчитали среднее значение \bar{y} . Реальные значения отклоняются от этой средней, можно измерить отклонение суммированием всех возведенных в квадрат погрешностей:

Вся SSE (сумма всех возведенных в квадрат погрешностей, или **СКП**) = $\sum(y_i - \bar{y})^2$

Полученное выражение суммы квадратических погрешностей (СКП), можно разделить на различные компоненты



Модель регрессии объясняет *некоторые* отклонения реальных наблюдений от средней:

$$\text{Объяснимая часть СКП} = \sum(\hat{y}_i - \bar{y})^2$$

Но есть еще то, что модель регрессии не объясняет всех отклонений, и кое-какие *остатки* так и останутся необъясненными:

$$\text{Необъяснимая часть СКП} = \sum(y_i - \bar{y})^2.$$

Найдем, что:

$$\text{Вся СКП} = \text{Объяснимые СКП} + \text{Необъяснимые СКП}.$$

Чем больше отклонение, объяснимое регрессией, тем точнее прямая наилучшего соответствия. Отсюда показатель наилучшего соответствия прямой данным — это отношение суммарных **СКП**, которое *объясняется* моделью регрессии. Это и есть **коэффициент детерминации**:

$$\text{коэффициент детерминации} = \frac{\text{Объяснимые СКП}}{\text{Вся СКП}}$$

Этот показатель принимает значения от 0 до 1. Если он близок к 1, тогда большая часть отклонений объяснена регрессией, необъяснимое отклонение невелико, и прямая идеально подходит к данным. Если значение около 0, то, наоборот, большая часть отклонений необъяснима, и прямая, хотя и лучший из вариантов, но тем не менее она все равно «низкого качества».

Математическое выражение для **коэффициента детерминации** следующее:

$$\text{коэффициента детерминации} = \left[\frac{n \sum xy - \sum x \sum y}{\sqrt{[n \sum x^2 - (\sum x)^2] * [n \sum y^2 - (\sum y)^2]}} \right]^2$$

Коэффициент детерминации обозначается r^2 или R^2 , и лучше всего доверить расчет компьютеру.

Обычно любое значение коэффициента детерминации свыше 0,5 считается хорошим. При более низком коэффициенте, скажем близком к 0,2, большая часть отклонения необъяснима с помощью регрессии, и зависимость несильная. Однако не следует забывать о единичных крайних, далеко отстоящих данных наблюдений. Даже случайное значение такого рода может сказаться на регрессии и снизить коэффициент детерминации, так что всегда есть искушение допустить, что они — ошибки, и проигнорировать их. Этого делать ни в коем случае нельзя! Такую точку можно отбросить, только если есть настоящая причина - ошибка или потому, что точка абсолютно не сопоставима с другими данными. Решение произвольно отбросить некоторые наблюдения, потому что они портят рисунок, приводят к потере самой цели анализа — оценить, есть ли некая линия и степень зависимости.

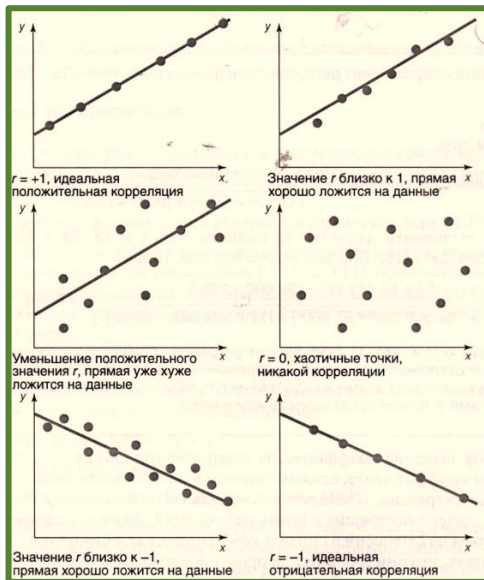
Коэффициент корреляции

Вторым показателем регрессии выступает **коэффициент корреляции**, который отвечает на основной вопрос: связаны ли x и y линейно? Коэффициенты корреляции и детерминации, очевидно, отвечают на очень схожие вопросы, и типичный результат показывает, что:

$$\text{Коэффициент корреляции} = \sqrt{\text{Коэффициент детерминации}}$$

Теперь понятно, почему мы обозначили коэффициент детерминации как r^2 - чтобы можно было обозначить коэффициент корреляции как r . Его называют еще коэффициентом Пирсона, и он принимает значения между +1 и -1.

Значение $r = 1$ показывает, что две переменные идеально связаны линейной зависимостью при полном отсутствии помех: когда одна растет, то же самое происходит и с другой. Интерпретация коэффициента корреляции показана на следующем рисунке:



При коэффициенте корреляции, близком к +1 или -1, между двумя переменными будет сильная зависимость. Однако, когда r падает до 0,7 или -0,7, коэффициент детерминации равен 0,49, т. е. становится очень низким. Можно заключить, что при значениях r от 0,7 до -0,7 зависимость достаточно слабая.

Пример – 15 наблюдений

Рассчитать коэффициенты детерминации и корреляции данных в таблице. Какой можно сделать из них вывод? Какова наилучшая прямая?

x	4	17	3	21	10	8	4	9	13	12	2	6	15	8	19
y	13	47	24	41	29	33	28	38	46	32	14	22	26	21	50

$$Y = 15,376 + 1,545x$$

$$r^2 = 0,6348$$

$$R = 0,7967$$

Ранговая корреляция

Коэффициент Пирсона — наиболее широко используемый показатель корреляции, но он работает только с количественными данными (числовыми значениями). Иногда нам необходимо замерить силу зависимости между порядковыми данными (данные ранжированы в определенном порядке, но их значения неизвестны). Представьте рыночное исследование, в котором людей просят указать порядок предпочтений по предложенным им альтернативам. Например, исследование может содержать предложение расставить всех интернет-провайдеров в порядке качества их услуг. Также можно предложить проранжировать цены за их услуги. Поэтому полезный анализ мог бы установить, есть ли связь между двумя списками, т. е. между качеством и ценой. Это можно сделать, используя коэффициент ранговой корреляции Спирмена, обозначаемый как r_s .

$$\text{Коэффициент Спирмена} = r_s = 1 - \frac{6 \sum D^2}{n(n^2 - 1)}$$

где n — число парных наблюдений;

D — разница в порядковых номерах (рангах);

D — Место(значение ранга) в первом списке – Место(значение ранга) во втором списке.

Пример – 5 услуг.

Услуга	V	И/	X	Y	Z
Ранжирование по качеству	2	5	13	3	4
Ранжирование по цене	1	3	2	4	5

Решение

По формуле Спирмена:

D = Место в списке качества - Место в списке цен.

В этом случае 5 мест, значит, $n = 5$ и сумма D^2 такова:
 $(2 - 1)^2 + (5 - 3)^2 + (1 - 2)^2 + (3 - 4)^2 + (4 - 5)^2 = 1 + 4 + 1 + 1 + 1 = 8$.

Коэффициент Спирмена: $r_s = 1 - \frac{6 \sum D^2}{n(n^2-1)} = 1 - \frac{6 \cdot 8}{5 \cdot (25-1)} = 0,6$

Хотя он выглядит совершенно отличным от коэффициента Пирсона, тем не менее коэффициент Спирмена использует тот же принцип и интерпретируется совершенно аналогично. То есть значение 0,6 говорит о том, что между ценой и качеством есть некоторая зависимость, но она не очень сильная.

Стоит помнить, что порядковые данные гораздо менее точны, чем количественные. То есть место 1 может быть чуть лучше места 2 по качеству, но может быть и намного лучше. Отсюда и результаты регрессии более условны. Для повышения точности следует использовать метод парных сравнений, позволяющий указать небольшие отличия в сравнение. Составляется матрица всевозможных парных сравнений. Поясним на предыдущем примере. Исследуются 5 услуг - V, W, X, Y, Z. Составляем матрицу

	V	W	X	Y	Z
V	1	2/10	2/3	1/2	3/4
W		1	3/5	3/7	1/2
X			1	5/6	4/2
Y				1	4/5
Z					1

Значимость (важность) каждой пары указывается в виде отношения двух, целых чисел из заранее выбранного диапазона. В нашем примере 1...10. Заполняется только над диагональная часть. Под диагональная содержит обратные отношения. Получили так называемую матрицу предпочтений. Ранги (вместо порядковых номеров) вычисляются по следующим формулам:

$$r_{ij} = \frac{2P_{ij}}{P_{ij} + 1}$$

$$r_{ji} = \frac{2}{P_{ij} + 1}$$

Согласованность (компетентные участники) участников определяется коэффициентом конкордации W, изменяется от 0 до 1. Интерпретация как у коэффициента детерминации.

$$W = \frac{12 \sum_{i=1}^n \left(\sum_{r=1}^m r_{ik} - \frac{m(n+1)}{2} \right)^2}{m^2 n (n^2 - 1)}$$

n – количество сравниваемых целей (объектов);

m – количество сравниваемых показателей;

k – количество участников оценки.

Если W имеет низкое значение, можно определить коэффициенты парной корреляции

для каждой пары участников опроса по Спирмену и найти не компетентных участников.

Коэффициент по Спирмену = $1 - \frac{6 \sum_{i=1}^n (r_{i1} - r_{i2})^2}{n(n^2 - 1)}$ для первого и второго участника (эксперта).

Множественная регрессия

Есть ряд продолжений метода линейной регрессии, самое популярное из которых привязывает к зависимой переменной несколько независимых. Поясним: допустим, у вас есть объем продаж, который можно «привязать» к бюджету рекламы, цене, уровню безработицы, среднему доходу граждан, к продажам у конкурентов и т. д. Иными словами, зависимая переменная y задается не одним параметром, независимой переменной x , а целым рядом независимых переменных x_i . Запишем эту зависимость в таком виде:

$$y = a + b_1 x_1 + b_2 x_2 + b_3 x_3 + \dots$$

Поясним на словах:

Продажи = $a + b_1 * \text{Бюджет рекламы} + b_2 * \text{Цена} + b_3 * \text{Ставка безработицы} + b_4 * \text{Доход} + b_5 * \text{Продажи у конкурентов}$.

Добавляя новые независимые переменные, мы стремимся получить более точную модель. Тогда мы сможем установить, что реклама объясняет 60% отклонений в продажах, но с новым членом, ценой уже 75% отклонений будет объяснено, а при добавлении сюда еще и уровня безработицы мы получим еще больший показатель — 85% и т. д.

Поскольку мы ищем линейную зависимость между зависимой переменной и множеством независимых, мы на самом деле должны обозначить этот тип связи как множественную линейную регрессию, которую сокращенно называют «**множественная регрессия**».

Множественная регрессия — это тоже линия, которая наилучшим образом ложится на множество зависимых переменных.

Она определяет наилучшие значения для a и b_i в уравнении

$$y = a + b_1 x_1 + b_2 x_2 + b_3 x_3 + \dots$$

Множественную регрессию никогда не считают вручную, а пользуются специальными функциями в компьютерных программах.