

# Корреляционно-регрессионный анализ в Excel

Финансовый университет



Конспект лекции

Смирнов Михил Викторович

Москва - 2021

# 1 Спецификация модели

Различают простую (парную) и множественную регрессию.

- Парная регрессия

$$\hat{y} = f(x)$$

- Множественная регрессия

$$\hat{y} = f(x_1, x_2, \dots, x_n)$$

$\hat{y}$  - объясняемая переменная Предпосылка моделирования - гипотеза о влиянии тех или иных факторов на объясняемую переменную. Пусть спрос  $y$  на товар находится в обратной зависимости от цены  $x$ :

$$\hat{y} = a - b \cdot x$$

Связь проявляется как закономерность в среднем по совокупности наблюдений. Например,  $y = 500 - 2 \cdot x$  означает, что с ростом цены на 1 д.е. спрос **в среднем** уменьшается на 2 единицы.

Каждое фактическое значение можно выразить через теоретическое значение и случайную величину отклонения, найденную по уравнению регрессии

$$y_j = \hat{y}_j + \epsilon_j$$

Величина  $\epsilon$  обусловлена факторами:

- спецификация модели;
- выборочный характер данных;
- ошибки измерения.

Тогда можем записать  $y = 500 - 2 \cdot x + \epsilon$ . Функция может быть нелинейной:

- $\hat{y} = a \cdot x^{-b}$ ;
- $\hat{y} = a + \frac{b}{x}$ ;
- $\hat{y} = \frac{1}{a+b \cdot x}$

**Ошибки спецификации модели.** Величина ошибки тем меньше, чем качественнее спецификация модели. К другим ошибкам спецификации относится неучет каких-либо дополнительных факторов.

**Ошибки выборки.** Происходят из-за неоднородности данных, из-за наличия выбросов и пропусков, а также из-за недостаточного объёма выборки. Выбор конкретного временного интервала, в котором измерены показатели, также влияет на результаты регрессии.

**Ошибки измерения.** На практике представляют наибольшую опасность. Например, реальные запасы полезных ископаемых могут существенно отличаться от их оценки из-за неверной интерпретации сопутствующих месторождению признаков.

**Линейные и нелинейные модели.** В большинстве случаев (в социально-экономических системах) линейные модели предпочтительнее нелинейных. Например, ввиду ограниченности выборки данных, нелинейность может не проявиться. Поэтому, вместо  $y = a + b \cdot x + c \cdot x^2 + \epsilon$  вполне достаточно использовать  $y = a + b \cdot x$ . Кроме того, при грубых измерениях, линейные модели менее чувствительны к ошибкам измерения.

**Выбор вида математической функции.** Выбор вида функции может осуществляться методами:

- графическим;
- аналитическим;
- экспериментальным.

**Графический метод** нагляден. На рис. 1 представлены основные виды используемых в ходе количественной оценки связей.

**Аналитический метод** основан на изучении природы связи исследуемых признаков. Допустим, изучается потребность предприятия в электроэнергии в зависимости от объема продукции. Потребление  $y$  можно разделить на потребление связанное с выпуском продукции на прямую ( $b \cdot x$ ), и связанную с организацией деятельности предприятия в целом ( $a$ ), например, отопление, водоснабжение и т.д. Тогда запишем уравнение регрессии

$$\hat{y} = a + b \cdot x$$

Разделив обе части уравнения на  $x$  получим выражение зависимости удельного расхода электроэнергии  $z = \frac{\hat{y}}{x}$  на единицу продукции от объема выпущенной продукции в форме гиперболы:

$$\hat{z} = b + \frac{a}{x}$$

В Excel выбор конкретной функции можно осуществлять с помощью величины остаточной дисперсии  $D_\epsilon$ , которая могла бы быть равной нулю в случае, если  $y = \hat{y}$ . Однако, в реальной жизни такого не происходит и остаточная дисперсия всегда положительна. Поэтому одним из оснований выбора функции является величина остаточной дисперсии, которая тем меньше, чем ближе теоретические данные к фактическим.

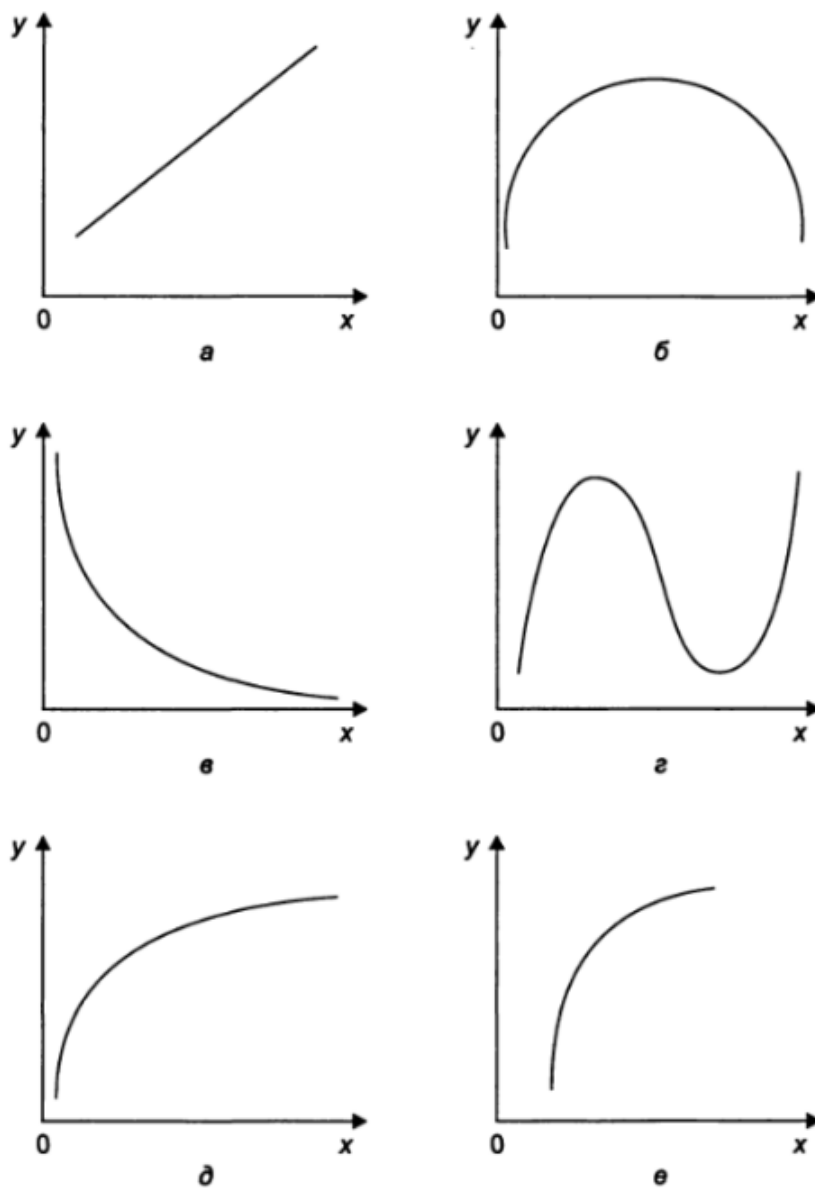
$$D_\epsilon = \frac{1}{M} \sum_{m=1}^M (y_i - \hat{y}_i)^2$$

Если  $D_\epsilon$  примерно одинаково для разных функций, то выбирают по-возможности, более простую функцию.

Опытным путем установлено, что число наблюдений должно в 6-7 раз превышать число рассчитываемых параметров при переменных. Так, например, для параболы второй степени

$$\hat{y} = a + b \cdot x + c \cdot x^2$$

требуется не менее 14 наблюдений. При прочих равных условиях предпочтительна модель с меньшим числом параметров.



**Рис. 1.** Основные типы кривых, используемые при количественной оценке связей между двумя переменными:

$$а - \hat{y}_x = a + b \cdot x;$$

$$в - \hat{y}_x = a + b/x;$$

$$д - \hat{y}_x = a \cdot x^b;$$

$$б - \hat{y}_x = a + b \cdot x + c \cdot x^2;$$

$$г - \hat{y}_x = a + b \cdot x + c \cdot x^2 + d \cdot x^3;$$

$$е - \hat{y}_x = a \cdot b^x$$

## 2 Линейная регрессия и корреляция

### Парная линейная регрессия

Парная линейная регрессия сводится к нахождению параметров  $a$  и  $b$  уравнения вида

$$\hat{y} = a + b \cdot x$$

или

$$y = a + b \cdot x + \epsilon$$

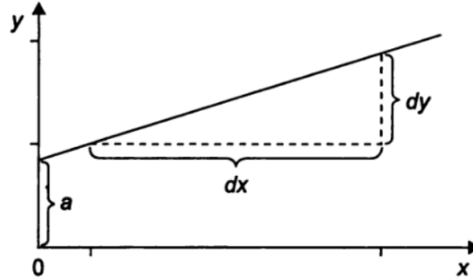


Рис. 2. Графическая оценка параметров линейной регрессии

Параметр  $a$  определим как точку пересечения линии регрессии с осью  $Oy$ , а параметр  $b$  — исходя из угла наклона линии регрессии как  $dy/dx$ , где  $dy$  — приращение результата  $y$ , а  $dx$  — приращение фактора  $x$ .

Классическим подходом к оцениванию  $a$  и  $b$  является использование метода наименьших квадратов (МНК), позволяющего получить такие значения  $a, b$ , при которых сумма квадратов разностей фактических и теоретических значений минимальна

$$S(a, b) = \sum_{m=1}^M (y_m - \hat{y}_m)^2 = \sum_{m=1}^M (y_m - a - b \cdot x_m)^2 \rightarrow \min$$

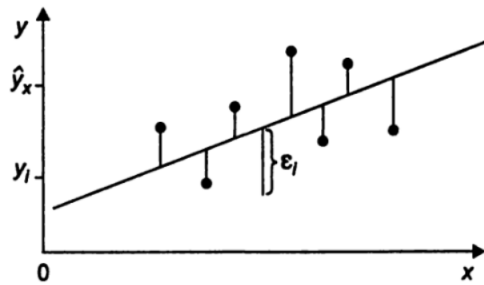


Рис. 3. Линия регрессии с минимальной дисперсией остатков

Чтобы найти минимум функции  $S(a, b)$  надо вычислить частные производные по каждому из параметров  $a$  и  $b$  и приравнять их к нулю. Решение системы уравнений приводит к результату:

$$a = \bar{y} - b \cdot \bar{x} \quad (1)$$

$$b = \frac{\bar{x}\bar{y} - \bar{x} \cdot \bar{y}}{\bar{x}^2 - (\bar{x})^2} \quad (2)$$

Докажем это.

$$\frac{dS}{da} = -2 \sum_{m=1}^M (y_m - a - b \cdot x_m) = 0$$

$$\frac{ds}{db} = -2 \sum_{m=1}^M (y_m - a - b \cdot x_m) \cdot x_m = 0$$

Получена система из двух уравнений с двумя неизвестными. Сократим на два и раскроем скобки.

$$\begin{aligned} M \cdot a + b \cdot \sum_{m=1}^M x_m &= \sum_{m=1}^M y_m \\ a \cdot \sum_{m=1}^M x_m + b \cdot \sum_{m=1}^M x_m^2 &= \sum_{m=1}^M y_m x_m \end{aligned}$$

Разделим оба уравнения на  $M$ .

$$\begin{aligned} a + b \cdot \bar{x} &= \bar{y} \\ a \cdot \bar{x} + b \cdot \bar{x}^2 &= \bar{x}\bar{y} \end{aligned}$$

В первом уравнении выразим  $a$  через  $b$

$$a = \bar{y} - b \cdot \bar{x}$$

Подставим  $a$  во второе уравнение

$$\bar{x} \cdot \bar{y} - b \cdot (\bar{x})^2 + b \cdot \bar{x}^2 = \bar{x}\bar{y}$$

Приведем подобные члены

$$b \cdot (\bar{x}^2 - (\bar{x})^2) = \bar{x}\bar{y} - \bar{x} \cdot \bar{y}$$

Оставим  $b$  в левой части уравнения, иксы перенесем направо

$$b = \frac{\bar{x}\bar{y} - \bar{x} \cdot \bar{y}}{\bar{x}^2 - (\bar{x})^2}$$

Можно показать, что  $\bar{x}\bar{y} - \bar{x} \cdot \bar{y} = Cov_{xy}$ , а  $\bar{x}^2 - (\bar{x})^2 = D_x$ , тогда

$$b = \frac{Cov_{xy}}{D_x}$$

Параметр  $b$  называется коэффициентом регрессии. Знак  $b$  показывает направление связи: при  $b > 0$  связь прямая, при  $b < 0$  связь обратная.

Параметр  $a$  может не иметь социально-экономического содержания, особенно, если  $a \leq 0$ .

### 3 Задача построения регрессионной модели в Excel

Рассмотрим задачу построения регрессионной модели в Excel. Воспользуемся данными о цене акций Газпрома и Сбербанка за период с 1 августа по 1 октября 2021 года. Источник данных - [finam.ru](https://www.finam.ru)

The screenshot shows the 'Finam.ru' website interface for exporting stock data. The main heading is 'МосБиржа акции' (Moscow Exchange Stocks) with a dropdown menu showing 'Сбербанк' (Sberbank). The form includes the following fields and options:

- Интервал и периодичность** (Interval and frequency): Date range from 01.08.2021 to 01.10.2021, frequency set to 1 день (1 day).
- Имя выходного файла** (Output file name): SBER\_210801\_211001, format set to .CSV.
- Имя контракта** (Contract name): SBER.
- Формат** (Format): даты (dates) set to dd/mm/yy, времени (time) set to ЧЧММСС.
- Выдавать время** (Output time): Radio buttons for 'начала свечи' (start of candle) and 'окончания свечи' (end of candle), with 'московское' (Moscow) checked.
- Разделитель** (Delimiter): 'полей' (fields) set to 'точка с запятой (;)' (semicolon), 'разрядов' (digits) set to 'нет' (none).
- Формат записи в файл** (File record format): TICKER, PER, DATE, TIME, OPEN, HIGH, LOW, CLOSE, VOL.
- Добавить заголовок файла** (Add file header): checked.
- Заполнять периоды без сделок** (Fill periods without trades): unchecked.

A sidebar on the left contains navigation links: Обзор new, Обзор old, О компании, Новости, Комментарии, Календарь событий, Календарь статистики, Корпоративные события, Теханализ Live! new, and Теханализ Live!.

Рис. 4. Форма запроса на получение данных