

ЭКОНОМЕТРИКА

ЛАБОРАТОРНАЯ РАБОТА №2

АНАЛИЗ КАЧЕСТВА РЕГРЕССИИ И ПРОГНОЗИРОВАНИЕ

Для выполнения расчётов в данной работе используются три способа:

- формулы
- функции
- надстройка

Отчёт по лабораторной работе оформляется в соответствии с требованиями ГОСТ ЕСКД (рамка и штамп не требуются).

Литература

1. Методические рекомендации по нормоконтролю
2. Методические указания по оформлению курсовых и дипломных работ
3. ГОСТ 2.105-95 ЕСКД. Общие требования к текстовым документам

Режим доступа к локальным нормативным документам:

Портал кафедры -> Документы -> Диплом -> Нормоконтроль

1. Оценивание уравнения регрессии

Регрессия – это «средняя» зависимость между независимой (объясняющей) переменной x и зависимой переменной y . Зависимость проявляется «в среднем», на фоне случайного разброса вокруг линии регрессии:

Уравнение регрессии выглядит следующим образом:

$$y = a + b \cdot x + e,$$

где a – свободный член уравнения;

b – коэффициент регрессии;

e – случайный член (остатки);

y – зависимая переменная, результативный признак;

x – независимая (объясняющая) переменная, факторный признак.

Результаты регрессионного анализа по методу наименьших квадратов (МНК) представляют в следующем виде:

$$y = \underset{(\sigma_a)}{\hat{a}} + \underset{(\sigma_b)}{\hat{b}} x$$

где σ_a и σ_b – стандартные отклонения коэффициентов уравнения регрессии.

Например:

$$y = 27,4 + 2,35 x$$

(2,51) (0,369)

Значения оценок коэффициентов и их стандартных отклонений записывают с учетом числа значащих цифр (разрядов). Например, во всех следующих числах по 3 значащих разряда:

234 000	234	23,4	2,34
0,234	0,000234	2,34 x 10⁵	23,4 x 10⁻³

Рис.1. Примеры чисел с тремя значащими разрядами

Используйте следующее правило: число значащих разрядов при записи результатов должно быть не меньше, чем в исходных данных.

В данной работе используются три способа регрессионного анализа.

Первый способ – надстройка «Анализ данных», см. рис.2.

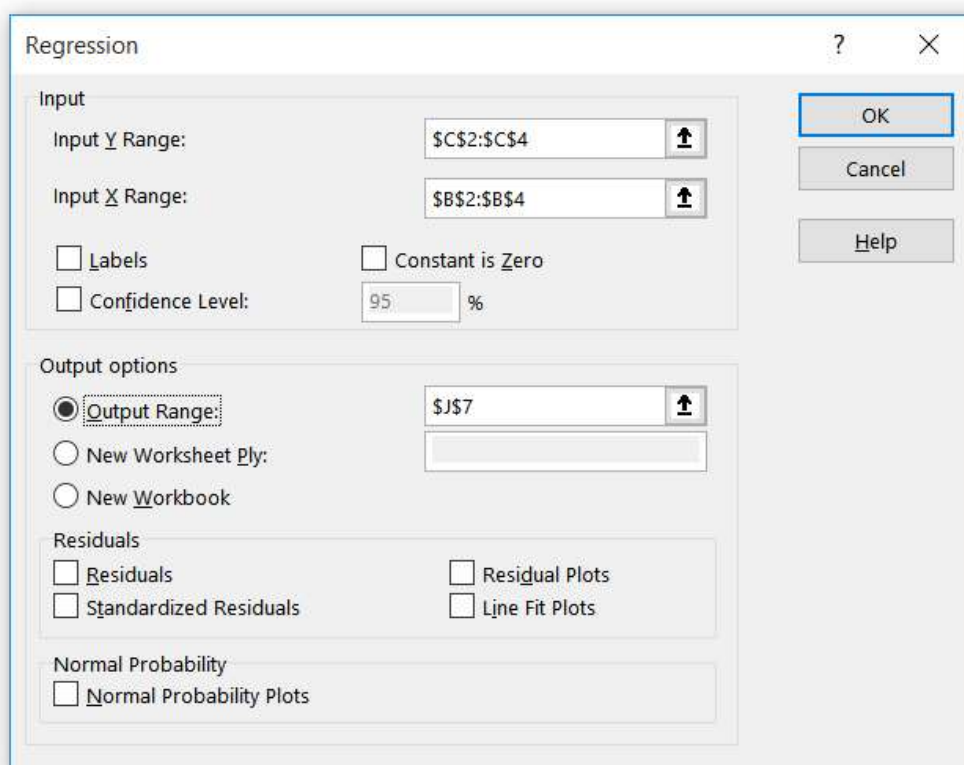


Рис.2. Регрессионный анализ в надстройке «Анализ данных»,

Второй способ – функция LINEST, ЛИНЕЙН, см. рис.3. Введите в выбранную ячейку имя функции, укажите параметры и нажмите [Enter]. Затем выделите диапазон ячеек, с начальной ячейкой в левом верхнем углу диапазона, нажмите [F2], а затем [Ctrl+Shift+Enter]. В диапазоне ячеек будут выведены оценки коэффициентов, их стандартные отклонения и другие показатели.

fx {=LINEST(C2:C4,B2:B4,1,1)}		
B	C	D
	0.626649077	-26.3878628
	0.262778157	48.58625288
	0.850452318	4.176988096
	5.686830498	1
	99.21943712	17.44722955

Рис.3. Функция LINEST

Третий способ – формулы. Для построения уравнения регрессии нужно решить систему нормальных уравнений:

$$\begin{cases} \Sigma y = an + b\Sigma x \\ \Sigma yx = a\Sigma x + b\Sigma x^2 \end{cases}$$

Решив данную систему, можно определить коэффициенты уравнения регрессии.

Чтобы записать систему уравнений, необходимо вначале рассчитать вспомогательные суммы. Расчёты проводятся по столбцам, внизу каждого столбца подсчитывают сумму.

Задание

1. Сгенерируйте исходные данные в соответствии с заданием к лабораторной работе по эконометрике №1.
2. Оцените параметры уравнения регрессии с помощью надстройки «Анализ данных».

3. Запишите уравнение регрессии с учётом числа значащих разрядов, указав оценки коэффициентов и их стандартные отклонения.

4. Изучите описание функции LINEST, ЛИНЕЙН.

5. Оцените параметры уравнения регрессии с помощью функции ЛИНЕЙН.

6. Сравните полученные результаты.

2. Коэффициент корреляции

Коэффициент линейной корреляции определяет тесноту линейно взаимосвязи между переменными. Его значение определяют по формуле:

$$r_{yx} = \frac{n \sum y \cdot x - \sum y \sum x}{\sqrt{\sum (n \sum x^2 - (\sum x)^2) (n \sum y^2 - (\sum y)^2)}}$$

Расчеты выполняются в пакете Excel. Исходные данные и прогнозы располагают в столбцах электронной таблицы.

Существует как минимум три способа выполнения расчётов в электронных таблицах:

- формула
- функция
- надстройка

Первый способ – формула, вводимая в ячейки. Для проведения расчётов необходимо построить вспомогательные столбцы произведений и квадратов величин, а затем найти суммы по каждому столбцу, см. рис.4.

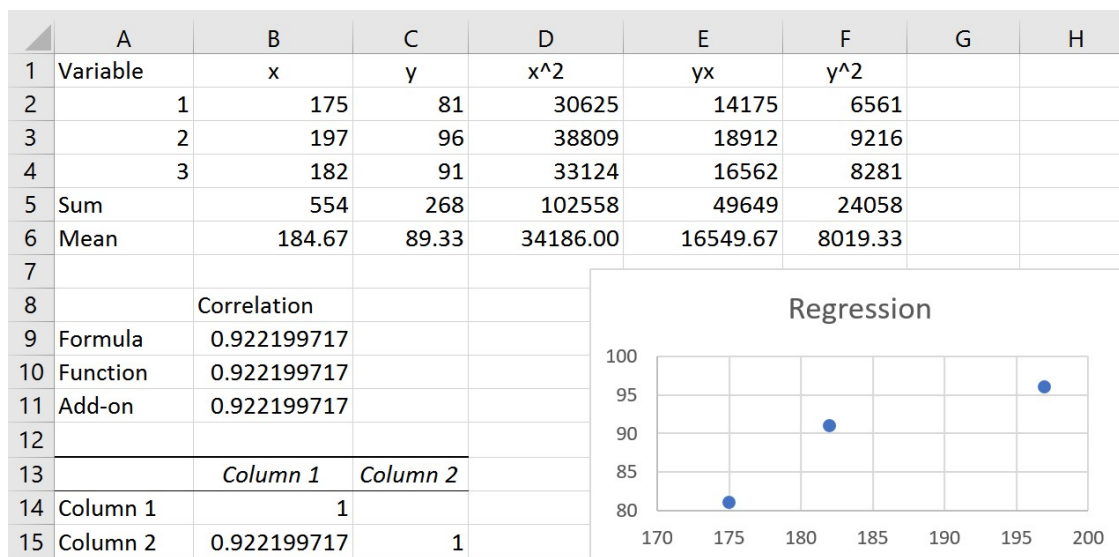


Рис.4. Вычисление коэффициента корреляции

Второй способ – использование готовых функций Excel. Для вычисления r_{yx} используется функция CORREL, см. рис.5.

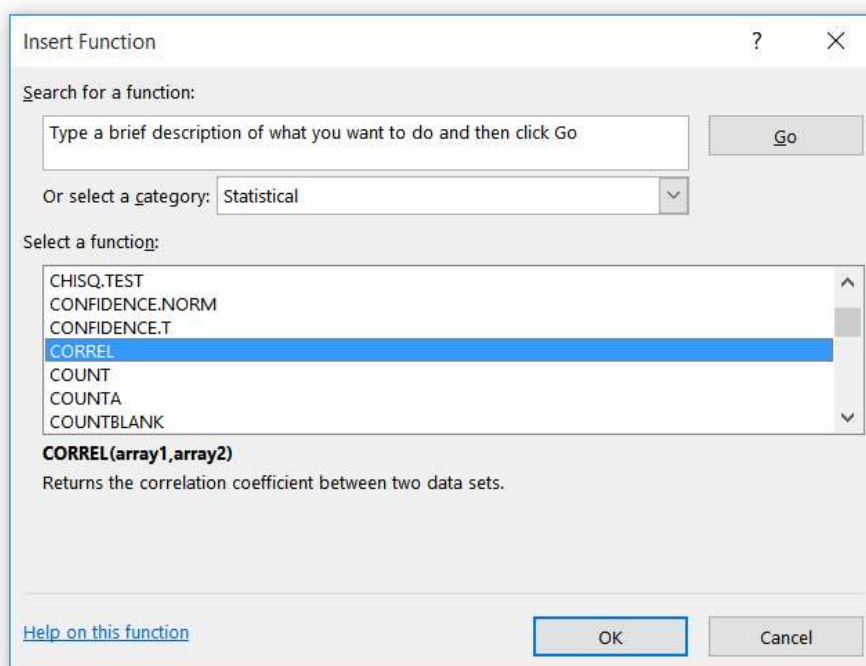


Рис.5. Статистическая функция CORREL

Третий способ – использование надстройки «Анализ данных», см. рис.6.

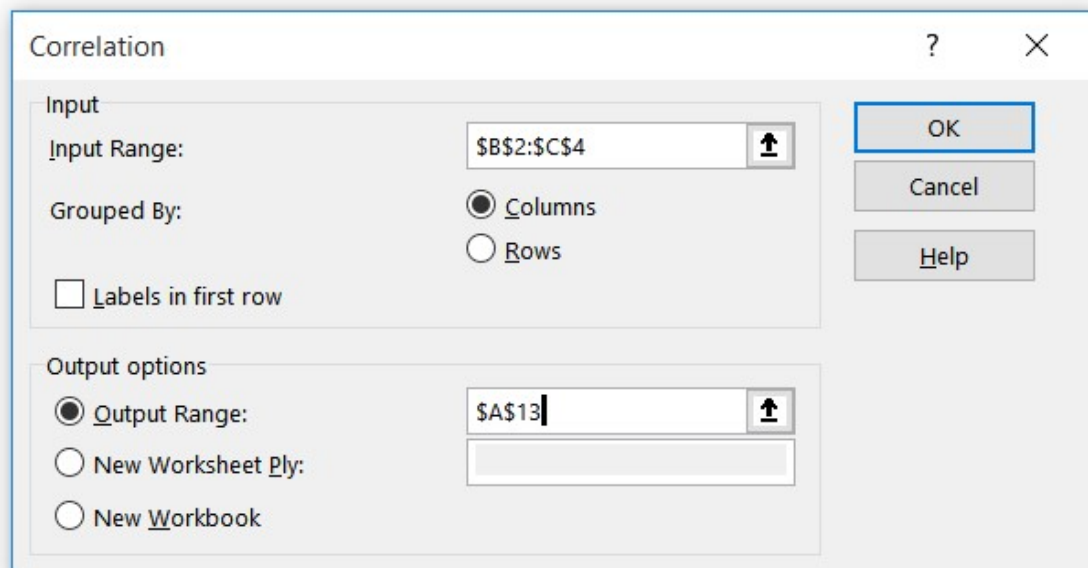


Рис.6. Вычисление коэффициента корреляции в надстройке

Кроме того, коэффициент корреляции можно вычислить следующим образом:

$$r_{yx} = \frac{\overline{y\bar{x}} - \bar{y} \cdot \bar{x}}{\sigma_y \cdot \sigma_x}$$

Вычисления любым способом должны давать близкие результаты (с точностью до погрешностей округления).

Задание

1. Определите r_{yx} всеми возможными способами.
2. Сравните результаты расчётов.
3. Сделайте вывод о степени тесноты линейной связи.

3. Коэффициент детерминации

Качество модели в целом характеризуется *коэффициентом детерминации* R^2 . Этот показатель представляет собой отношение объясненной и полной суммы квадратов результативного признака:

$$R^2 = \frac{ESS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

где ESS (Explained Sum of Squares) – объяснённая сумма квадратов;

TSS (Total Sum of Squares) – полная (общая) сумма квадратов;

RSS (Residual Sum of Squares) – сумма квадратов остатков.

TSS , ESS и RSS связаны следующим образом:

$$TSS = ESS + RSS$$

Формула для вычисления коэффициента детерминации:

$$R^2 = \frac{\sum(\hat{y}(x_i) - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

где $\{x_i, y_i\}$ – исходные данные;

$\bar{y}$ – среднее значение y ;

$\hat{y}(x_i)$ – оценка результативного признака при заданном значении факторного признака, линейный прогноз по уравнению регрессии.

Для парной линейной регрессии коэффициент детерминации равен квадрату коэффициента линейной корреляции:

$$R^2 = r^2$$

Для учёта сложности модели вводится *скорректированный* (adjusted) *коэффициент детерминации*:

$$R_{adj}^2 = 1 - \frac{RSS/(n - k)}{TSS/(n - 1)}$$

где k – количество коэффициентов уравнения регрессии;

n – объём выборки.

Скорректированный R^2 используют для сравнения разных моделей.

Задание

1. Определите $y(x_i)$ по уравнению регрессии для каждого x_i .
2. Вычислите TSS , ESS и RSS
3. Вычислите R^2 всеми возможными способами
4. Найдите r^2 и сравните с R^2 .
6. Сравните R^2 с результатами регрессионного анализа, полученными с помощью надстройки.
7. Определите скорректированный R^2 .

4. Качество уравнения регрессии

Статистическая значимость уравнения регрессии в целом проверяется путем построения F -статистики:

$$F = \frac{\sum (\hat{y}(x) - \bar{y})^2}{\sum (y - \hat{y}(x))^2 / (n - 2)} = \frac{R^2}{1 - R^2} \cdot (n - 2).$$

Фактическая статистика сравнивается с табличным значением для $df=\{1, n-2\}$ степеней свободы (Degrees of Freedom) и уровня значимости α . Если $F_{\text{факт}} > F_{1, (n-2), \alpha}$, то гипотезу о несущественности уравнения регрессии ($H_0: b=0$) отклоняют.

В некоторых случаях Excel позволяет дополнительно определить уровень значимости, при котором отвергается нулевая гипотеза (p-level).

Задание

1. Определите фактическое и критическое значения F-статистики
2. Проверьте значимость уравнения регрессии

5. Доверительные интервалы для коэффициентов

Доверительные интервалы для коэффициентов имеют вид:

$$\left| \frac{a - \hat{a}}{\sigma_a} \right| \leq t \left(\frac{1+p}{2}, n-2 \right)$$

$$\hat{a} - t \cdot \sigma_a \leq a \leq \hat{a} + t \cdot \sigma_a,$$

где a - неизвестное истинное значение коэффициента,

$\hat{a}$ – оценка,

t – квантиль распределения Стьюдента (критическое значение),

p – доверительная вероятность.

Интервал для коэффициента b строится аналогично.

Дисперсия оценок коэффициентов составляет:

$$\sigma_a^2 = \frac{\sum (y - \hat{y}(x))^2 \cdot \sum x^2}{(n-2) \cdot n \cdot \sum (x - \bar{x})^2},$$

$$\sigma_b^2 = \frac{\sum (y - \hat{y}(x))^2}{(n-2) \cdot \sum (x - \bar{x})^2}.$$

Для упрощения вычислений можно использовать приближённые «стандартные» значения коэффициента доверия t :

$$t(n \rightarrow \infty, p = 68\%) \approx 1$$

$$t(n \rightarrow \infty, p = 95\%) \approx 2$$

$$t(n \rightarrow \infty, p = 99,7\%) \approx 3$$

Для более точных расчётов используют квантили стандартного распределения Стьюдента.

Задание

1. Постройте доверительные интервалы для коэффициентов упрощённым способом.
2. Постройте доверительные интервалы для коэффициентов с использованием квантилей распределения Стьюдента
3. Сравните полученные интервалы

6. Значимость коэффициентов

Значимость коэффициентов проверяется по t -статистике Стьюдента:

$$t_a = \left| \frac{a}{\sigma_a} \right|, \quad t_b = \left| \frac{b}{\sigma_b} \right|.$$

Фактические значения сравниваются с табличными (критическими) для $df=(n-2)$ степеней свободы и уровня значимости α . Если $t_{\text{факт}} > t_{(n-2), \alpha}$, то гипотезу о несущественности коэффициента отклоняют.

Проверка значимости коэффициентов уравнения представляет собой исследование возможности равенства этих параметров нулю. Наиболее

простая трактовка статистической значимости такова. Если ноль находится внутри доверительного интервала, то оценка может принять нулевое значение. При этом говорят, что оценка статистически незначима. Если же оценка не может принять нулевое значение при заданной доверительной вероятности, то говорят, что оценка статистически значима, то есть *значимо отличается от нуля*.

Задание

1. Проверьте значимость коэффициентов упрощённым способом.
2. Проверьте значимость коэффициентов с использованием квантилей распределения Стьюдента
3. Сравните полученные результаты

7. Прогнозирование

Теоретические значения результативного признака (линейный прогноз) строятся на основе оценок параметров уравнения регрессии по имеющимся значениям факторного признака:

$$\hat{y}(x_i) = \hat{a} + \hat{b}x_i$$

Можно также сгенерировать столбец значений x от минимального до максимального значения с некоторым постоянным шагом.

Кроме самого значения определяют диапазон значений (доверительный интервал) при заданной вероятности, например при $p = 99,7 \%$.

Прогноз значений зависимой переменной строится в двух вариантах:

- прогноз средних значений

- прогноз индивидуальных значений.

Доверительные интервалы в этих случаях различаются.

Во втором случае дополнительно учитывается остаточная дисперсия (дисперсия остатков):

$$\sigma_e^2 = \frac{\sum (y - \hat{y}(x))^2}{n - 2} .$$

Задание

1. Постройте прогнозы для среднего и индивидуальных значений.
2. Опишите результаты в отчёте.