Пакет анализа «Регрессия»

Теория и практические советы

Про корреляцию & регрессию

- Задача корреляционного анализа – определение тесноты и направления связи между изучаемыми величинами.
- В ходе регрессионного анализа определяется аналитическое выражение связи зависимой случайной величины Y (результативный признак) с независимыми случайными величинами $X_1, X_2, \ldots X_m$ (факторами).

Уравнение регрессии -

это форма связи результативного признака Y с факторами X1, X2, …Xm. В зависимости от типа выбранного уравнения различают линейную и нелинейную (квадратичную, экспоненциальную, логарифмическую и т. д.) регрессию.

Парная и множественная

- В зависимости от числа взаимосвязанных признаков различают парную и множественную регрессию.
- Парная – исследуется связь между двумя признаками (результативным и факторным).
- Множественная (многофакторная) – между тремя признаками (результативным и несколькими факторными).

Задачи регрессионного анализа

- При помощи регрессионного анализа возможно решение задачи прогнозирования. Прогнозные значения вычисляются путем подстановки в уравнение регрессии параметров значений объясняющих переменных.
- **Задачи регрессионного анализа**
- Рассмотрим основные задачи регрессионного анализа: установление формы зависимости, определение *функции регрессии*, оценка неизвестных значений зависимой переменной.
- **Установление формы зависимости**.
- Характер и форма зависимости между переменными могут образовывать следующие разновидности регрессии:
- положительная линейная регрессия (выражается в равномерном росте функции);
- положительная равноускоренно возрастающая регрессия;
- положительная равнозамедленно возрастающая регрессия;
- отрицательная линейная регрессия (выражается в равномерном падении функции);
- отрицательная равноускоренно убывающая регрессия;
- отрицательная равнозамедленно убывающая регрессия.
- Однако описанные разновидности обычно встречаются не в чистом виде, а в сочетании друг с другом. В таком случае говорят о комбинированных формах регрессии.

Особенность и этапы регрессионного анализа

- Основная особенность регрессионного анализа: при его помощи можно получить конкретные сведения о том, какую форму и характер имеет зависимость между исследуемыми переменными.
- **Последовательность этапов регрессионного анализа**
- Формулировка задачи. На этом этапе формируются предварительные гипотезы о зависимости исследуемых явлений.
- Определение зависимых и независимых (объясняющих) переменных.
- Сбор статистических данных. Данные должны быть собраны для каждой из переменных, включенных в регрессионную модель.
- Формулировка гипотезы о форме связи (парная или множественная, линейная или нелинейная).
- Определение ***функции регрессии*** (заключается в расчете численных значений параметров уравнения регрессии)
- Оценка точности регрессионного анализа.
- Интерпретация полученных результатов. Полученные результаты регрессионного анализа сравниваются с предварительными гипотезами. Оценивается корректность и правдоподобие полученных результатов.
- Предсказание неизвестных значений зависимой переменной.

Этапы регрессионного анализа

- 1. Задание аналитической формы уравнения регрессии и определение параметров регрессии.
- 2. Определение в регрессии степени стохастической взаимосвязи результативного признака и факторов, проверка общего качества уравнения регрессии.
- 3. Проверка статистической значимости каждого коэффициента уравнения регрессии и определение их доверительных интервалов.

Зачем

- Инструмент анализа "Регрессия" применяется для подбора параметров уравнения регрессии с помощью метода наименьших квадратов. Регрессия используется для анализа воздействия на отдельную зависимую переменную значений одной или нескольких независимых переменных. Например, на спортивные качества атлета влияют несколько факторов, включая возраст, рост и вес. Можно вычислить степень влияния каждого из этих трех факторов по результатам выступления спортсмена, а затем использовать полученные данные для предсказания выступления другого спортсмена.
- Инструмент "Регрессия" использует функцию **ЛИНЕЙН**.

Определение функции и оценка неизвестных значений

- **Определение функции регрессии**.
- Вторая задача сводится к выяснению действия на зависимую переменную главных факторов или причин, при неизменных прочих равных условиях, и при условии исключения воздействия на зависимую переменную случайных элементов. *Функция регрессии* определяется в виде математического уравнения того или иного типа.
- **Оценка неизвестных значений зависимой переменной**.
- Решение этой задачи сводится к решению задачи одного из типов:
- Оценка значений зависимой переменной внутри рассматриваемого интервала исходных данных, т.е. пропущенных значений; при этом решается задача интерполяции.
- Оценка будущих значений зависимой переменной, т.е. нахождение значений вне заданного интервала исходных данных; при этом решается задача экстраполяции.
- Обе задачи решаются путем подстановки в уравнение регрессии найденных оценок параметров значений независимых переменных. Результат решения уравнения представляет собой оценку значения целевой (зависимой) переменной.

Предположения РА

- Рассмотрим некоторые предположения, на которые опирается регрессионный анализ.
- Предположение линейности, т.е. предполагается, что связь между рассматриваемыми переменными является линейной. Так, в рассматриваемом примере мы построили диаграмму рассеивания и смогли увидеть явную линейную связь. Если же на диаграмме рассеивания переменных мы видим явное отсутствие линейной связи, т. е. присутствует нелинейная связь, следует использовать нелинейные методы анализа.
- Предположение о нормальности *остатков*. Оно допускает, что распределение разницы предсказанных и наблюдаемых значений является нормальным. Для визуального определения характера распределения можно воспользоваться гистограммами *остатков*.
- При использовании регрессионного анализа следует учитывать его основное ограничение. Оно состоит в том, что регрессионный анализ позволяет обнаружить лишь зависимости, а не связи, лежащие в основе этих зависимостей.
- Регрессионный анализ дает возможность оценить степень связи между переменными путем вычисления предполагаемого значения переменной на основании нескольких известных значений.

Уравнение регрессии

- Уравнение регрессии выглядит следующим образом: Y=a+b*X
- При помощи этого уравнения переменная Y выражается через константу a и угол наклона прямой (или угловой коэффициент) b, умноженный на значение переменной X. Константу a также называют свободным членом, а угловой коэффициент - коэффициентом регрессии или B-коэффициентом.
- В большинстве случав (если не всегда) наблюдается определенный разброс наблюдений относительно регрессионной прямой.
- ***Остаток*** - это отклонение отдельной точки (наблюдения) от линии регрессии (предсказанного значения).
- Для решения задачи регрессионного анализа в MS Excel выбираем в меню **Сервис "Пакет анализа"** и инструмент анализа "Регрессия". Задаем входные интервалы X и Y. Входной интервал Y - это диапазон зависимых анализируемых данных, он должен включать один столбец. Входной интервал X - это диапазон независимых данных, которые необходимо проанализировать. Число входных диапазонов должно быть не больше 16.
- На выходе процедуры в выходном диапазоне получаем отчет, приведенный в следующих таблицах.

Этап 1

- Уравнение множественной линейной регрессии

$$\hat{y} = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_m x_m,$$

где $\hat{y}$ – теоретические значения результативного признака, полученные путем подстановки соответствующих значений факторных признаков в уравнение регрессии;

$x_1, x_2, \ldots x_m$ – значения факторных признаков;

$a_0, a_1, \ldots a_m$ – параметры уравнения (коэффициенты регрессии).

МНК

- Параметры уравнения регрессии могут быть определены с помощью метода наименьших квадратов, который используется в пакете анализа данных «Регрессия»: находятся параметры модели , при которых минимизируется сумма квадратов отклонений эмпирических (фактических) значений результативного признака от теоретических, полученных по выбранному уравнению регрессии, т.е.

$$S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_{1i} - a_2 x_{2_i} - \dots - a_m x_{mi})^2$$

- $\rightarrow$ min.

МНК

- Рассматривая S в качестве функции параметров a_i и проводя математические преобразования (дифференцирование), получаем систему нормальных уравнений с m неизвестными (по числу параметров a_i).

$$\begin{cases} \sum y = na_0 + a_1 \sum x_1 + a_2 \sum x_2 + \ldots + a_m \sum x_m, \\ \sum yx_1 = a_0 \sum x_1 + a_1 \sum x_1^2 + a_2 \sum x_2 x_1 + \ldots + a_m \sum x_m x_1, \\ \ldots \\ \sum yx_m = a_0 \sum x_m + a_1 \sum x_1 x_m + a_2 \sum x_2 x_m + \ldots + a_m \sum x_m^2. \end{cases}$$

Здесь n – число наблюдений, m – число факторов в уравнении регрессии.

Решение системы позволяет получить значения параметров регрессии a_i.

Этап 2

- Для определения величины степени стохастической взаимосвязи результативного признака Y и факторов X необходимо знать следующие дисперсии:
- - общую дисперсию результативного признака Y, отображающую влияние как основных, так и остаточных факторов:

$$\sigma_y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n},$$

- где $\bar{y}$ - среднее значение результативного признака Y;

Дисперсии

- - факторную дисперсию результативного признака Y, отображающую влияние только основных факторов:

$$\sigma^2_{\hat{O}} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{n};$$

- - остаточную дисперсию результативного признака Y, отображающую влияние только остаточных факторов:

$$\sigma^2_{O} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - (m+1)}.$$

Сложение дисперсий

- При корреляционной связи результативного признака и факторов выполняется соотношение

$$\sigma^2_{\hat{O}} < \sigma^2_y,$$

при этом

$$\sigma^2_y = \sigma^2_{\hat{O}} + \sigma^2_O.$$

Коэффициент детерминации R^2

- Для анализа общего качества уравнения линейной многофакторной регрессии используют множественный коэффициент детерминации R^2,
- называемый также квадратом коэффициента множественной корреляции R. Множественный коэффициент детерминации рассчитывается по формуле

$$R^2 = \frac{\sigma^2_{\hat{o}}}{\sigma^2_y}$$

- и определяет долю вариации результативного признака, обусловленную изменением факторных признаков, входящих в многофакторную регрессионную модель.

- Величина *R-квадрат,* называемая также мерой определенности, характеризует качество полученной регрессионной прямой. Это качество выражается степенью соответствия между исходными данными и регрессионной моделью (расчетными данными). Мера определенности всегда находится в пределах интервала [0;1].
- В большинстве случаев значение *R-квадрат* находится между этими значениями, называемыми экстремальными, т.е. между нулем и единицей.
- Если значение *R-квадрата* близко к единице, это означает, что построенная модель объясняет почти всю изменчивость соответствующих переменных. И наоборот, значение *R-квадрата,* близкое к нулю, означает плохое качество построенной модели.
- ***множественный R*** - коэффициент множественной корреляции R - выражает степень зависимости независимых переменных (X) и зависимой переменной (Y).
- *Множественный R* равен квадратному корню из коэффициента детерминации, эта величина принимает значения в интервале от нуля до единицы.
- В простом линейном регрессионном анализе *множественный R* равен коэффициенту корреляции Пирсона.

F критерий

- Так как в большинстве случаев уравнение регрессии приходится строить на основе выборочных данных, то возникает вопрос об адекватности построенного уравнения данным генеральной совокупности. Для этого проводится проверка статистической значимости коэффициента детерминации R^2 на основе F-критерия Фишера:

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m},$$

- где n – число наблюдений;
- m – число факторов в уравнении регрессии.

Если в уравнении регрессии свободный член $a_0 = 0$, то числитель n-m-1 следует увеличить на 1, т.е. он будет равен n-m.

F критерий

- В математической статистике доказывается, что если гипотеза $H_0 : R^2 = 0$ выполняется, то величина F имеет F-распределение с

k=m и i=n-m-1 числом степеней свободы, т.е.

$$\frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m} = F(k=m, l=n-m-1).$$

$H_0 : R^2 = 0$

R^2 Гипотеза о незначимости коэффициента детерминации $F_p > F_{\text{табл},\alpha}$.

$R^2 > 0{,}7$ отвергается, если

При значениях считается, что вариация результативного признака Y обусловлена в основном влиянием включенных в регрессионную модель факторов X.

Ошибка аппроксимации

- Для оценки адекватности уравнения регрессии часто также используют показатель средней ошибки аппроксимации

$$\bar{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|y_i - \hat{y}\right|}{y_i}\cdot 100\%.$$

Этап 3

- Возможна ситуация, когда часть вычисленных коэффициентов регрессии не обладает необходимой степенью значимости, т.е. значения данных коэффициентов будут меньше их стандартной ошибки. В этом случае такие коэффициенты должны быть исключены из уравнения регрессии. Поэтому проверка адекватности построенного уравнения регрессии наряду с проверкой значимости R^2 коэффициента детерминации включает также и проверку значимости каждого коэффициента регрессии.

t-критерий

- Для оценки адекватности уравнения регрессии часто также используют показатель средней ошибки аппроксимации

$$t = \frac{a_i}{\sigma_{a_i}},$$

где σ_{a_i} - стандартное значение ошибки для коэффициента регрессии a_i.

t-критерий

- В математической статистике доказывается, что если гипотеза $H_0 : a_i = 0$ выполняется, то величина t имеет распределение Стьюдента с k=n-m-1 числом степеней свободы, т.е.

$$\frac{a_i}{\sigma_{a_i}} = t(k = n - m - 1).$$

- Гипотеза $H_0 : a_i = 0$ о незначимости коэффициента регрессии отвергается, если $\left|t_p\right| > \left|t_{\hat{e}p}\right|$.

Границы доверительных интервалов

- Зная значение $t_{\hat{e}p}$, можно найти границы доверительных интервалов для коэффициентов регрессии

$$a_i^{\min} = a_i - t_{\hat{e}\eth}\sigma_{a_i};$$

$$a_i^{\max} = a_i + t_{\hat{e}\eth}\sigma_{a_i}.$$

Коэффициент эластичности

- При экономической интерпретации уравнения регрессии также широко используются частные коэффициенты эластичности, показывающие , на сколько процентов в среднем изменится значение результативного признака при изменении значения соответствующего факторного признака на 1%, и определяемые по формуле

$$\acute{Y}_{\tilde{O}_i} = a_i \frac{\overline{x}_i}{\overline{y}},$$

где $\frac{\overline{x}_i}{\overline{y}}$ - среднее значение соответствующего факторного признака;

a_i - среднее значение результативного признака;

- коэффициент регрессии при соответствующем факторном признаке.

Технология работы

Режим работы «Регрессия» служит для расчета параметров уравнения *линейной* регрессии и проверки его адекватности исследуемому процессу.

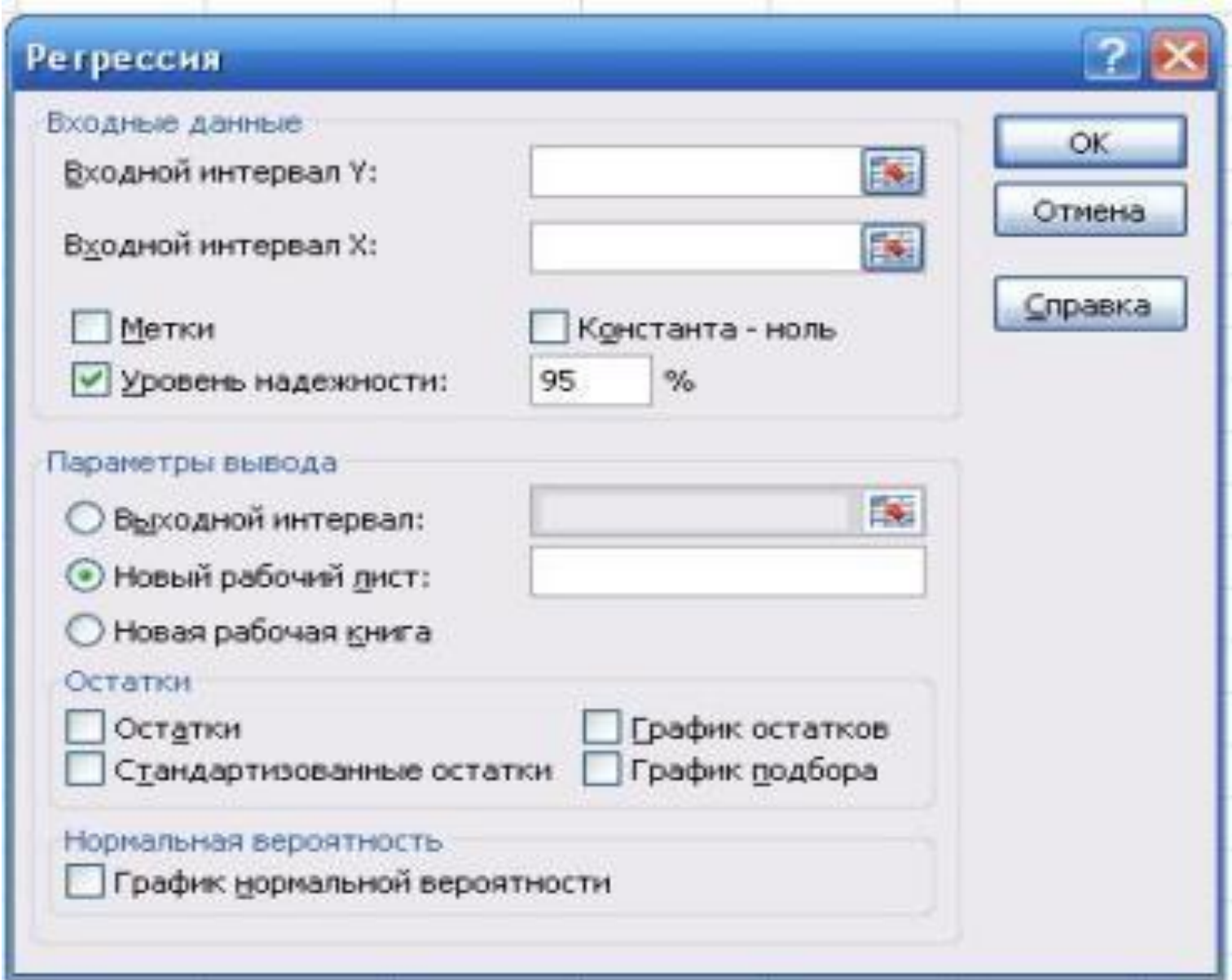
В диалоговом окне данного режима (рис. 14.1) задаются следующие параметры:

1. *Входной интервал Y* – вводится ссылка на ячейки, содержащие данные по результативному признаку. Диапазон должен состоять из одного столбца.

2. *Входной интервал X* – вводится ссылка на ячейки, содержащие факторные признаки. Максимальное число входных диапазонов (столбцов) равно 16.

3. Метки в первой строке/Метки в первом столбце. Флажок *Метки* устанавливается в активное состояние, если первая строка (столбец) во входном диапазоне содержит заголовки. Если заголовки отсутствуют, флажок следует деактивизировать. В этом случае будут созданы стандартные названия для данных выходного диапазона.

Рис. 14-1


Регрессия
Входные данные
Входной интервал Y:
Входной интервал X:
Метки
Константа - ноль
Уровень надежности:
95
%
ОК
Отмена
Справка
Параметры вывода
Выходной интервал:
Новый рабочий лист:
Новая рабочая книга
Остатки
Остатки
Стандартизованные остатки
График остатков
График подбора
Нормальная вероятность
График нормальной вероятности

Подготовка данных для ввода

- К сожалению, пакет анализа данных принимает в качестве входного интервала только данные, идущие подряд. Нельзя через точку с запятой перечислять массивы, находящиеся в разных местах файла. Кроме того, каждый показатель должен быть прописан по столбцам сверху вниз. Должно быть одинаковое количество значений в каждой вводимой переменной. Если необходимо вводить переменные текущего периода и лаговые, их следует выстроить на одинаковом уровне, а только подписать, где текущая переменная, а где лаговая.

  Пример входного массива дан на следующем слайде.

Пример массива, сформированного для ввода

1996	5537,495	3,658	3,79	3,614	71,486	82,557	5537,495	2
1997	6166,754	3,79	4,107	8,341	82,557	88,441	6166,754	3
1998	6600,589	4,107	4,303	4,795	88,441	92,279	6600,589	4
1999	6977,678	4,303	4,443	3,253	92,279	95,984	6977,678	5
2000	7691,83	4,443	4,751	6,916	95,984	100	7691,83	6
2001	8545,875	4,751	5,133	8,048	100	101,691	8545,875	7
2002	9319,317	5,133	5,465	6,474	101,691	105,359	9319,317	8
2003	10262,03	5,465	5,858	7,191	105,359	109,121	10262,03	9
2004	11505,78	5,858	6,367	8,675	109,121	116,772	11505,78	10
2005	13181,39	6,367	7,042	10,602	116,772	128,649	13181,39	11
2006	15117,33	7,042	7,783	10,527	128,649	142,961	15117,33	12
2007	17148,68	7,783	8,53	9,6	142,961	172,572	17148,68	13
2008	17032,59	8,53	8,251	-3,275	172,572	194,949	17032,59	14
2009	14220,84	8,251	6,788	-17,729	194,949	192,554	14220,84	15
2010	14418,72	6,788	6,765	-0,335	192,554	188,298	13000	16

4-5

4. *Уровень надежности* – установите данный флажок в активное состояние, если в поле, расположенное напротив флажка, необходимо ввести уровень надежности, отличный от уровня 95 %, применяемого по умолчанию. Установленный уровень надежности используется для проверки значимости коэффициента детерминации R^2 и коэффициентов регрессии a_i.

Примечание. При неактивном флажке *Уровень надежности* в таблице параметров уравнения регрессии (см. табл. 14.4, 14.9) генерируются две одинаковые пары столбцов для границ доверительных интервалов.

5. *Константа-ноль* – установите данный флажок в активное состояние, если требуется, чтобы линия регрессии прошла через начало координат (т. е. $a_0 = 0$).

6

6. Выходной интервал/Новый рабочий лист/Новая рабочая книга.

В положении *Выходной интервал* активизируется поле, в которое необходимо ввести ссылку на левую верхнюю ячейку выходного диапазона. Размер выходного диапазона будет определен автоматически, и на экране появится сообщение в случае возможного наложения выходного диапазона на исходные данные.

В положении *Новый рабочий лист* открывается новый лист, в который начиная с ячейки A1 вставляются результаты анализа. Если необходимо задать имя открываемого нового рабочего листа, введите его имя в поле, расположенное напротив соответствующего положения переключателя.

В положении *Новая рабочая книга* открывается новая книга, на первом листе которой начиная с ячейки A1 вставляются результаты анализа.

7-8

7. *Остатки* – установите данный флажок в активное состояние, если требуется включить в выходной диапазон столбец остатков (*см.* столбец *Остатки* в табл. 14.5).

8. *Стандартизованные остатки* – установите данный флажок в активное состояние, если требуется включить в выходной диапазон столбец стандартизованных остатков (*см.* столбец *Стандартизованные остатки* в табл. 14.5).

9-11

9. *График остатков* – установите данный флажок в активное состояние, если требуется вывести на рабочий лист точечные графики зависимости остатков от факторных признаков x_i.

10. *График подбора* – установите данный флажок в активное состояние, если требуется вывести на рабочий лист точечные графики зависимости теоретических результативных значений $\hat{y}$ от факторных признаков x_i.

11. *График нормальной вероятности* – установите данный флажок в активное состояние, если требуется вывести на рабочий лист точечный график зависимости наблюдаемых значений y от автоматически формируемых интервалов персентилей. График строится на основе генерируемой таблицы «Вывод вероятности» (*см.* табл. 14.6).

Пример 14.1

Пример 14.1. Данные о прибыли предприятий Y, величине оборотных средств X_1 и стоимости основных фондов X_2 приведены в табл. 14.1, сформированной на рабочем листе Microsoft Excel.

Табл. 14.1

	B	C	D	E
2	Номер предприятия	Прибыль Y, млн руб.	Величина оборотного капитала X1, млн руб.	Стоимость основного капитала X2, млн руб.
3	1	188	129	510
4	2	78	64	190
5	3	93	69	240
6	4	152	87	470
7	5	55	47	110
8	6	161	102	420

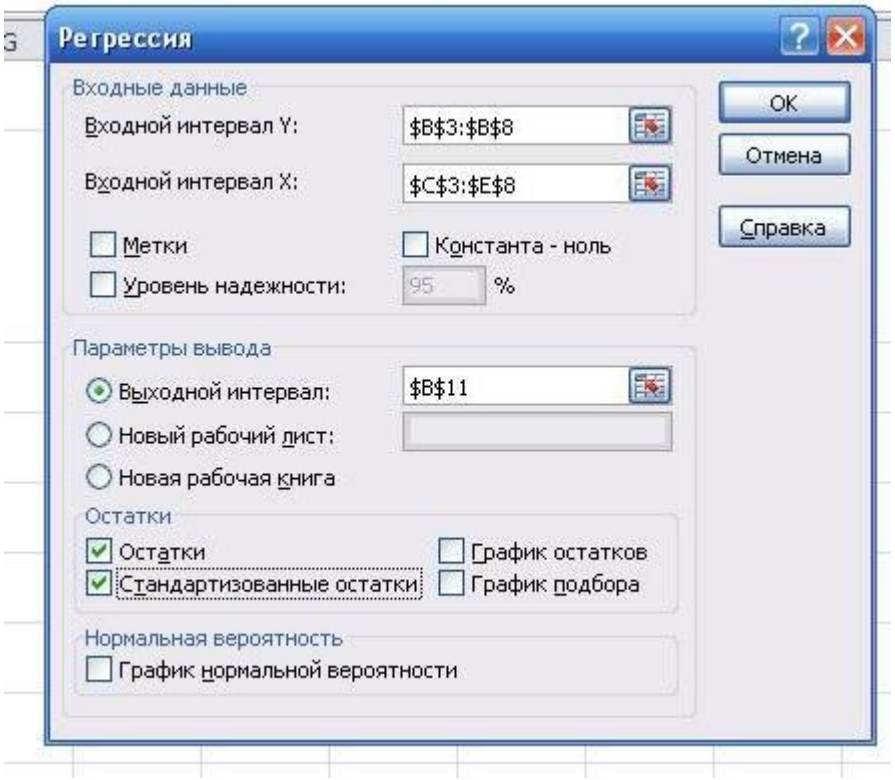
Что где

По представленным данным необходимо определить параметры уравнения линейной регрессии и провести его анализ.

Для решения задачи используем режим работы «Регрессия». Значения параметров, установленных в одноименном диалоговом окне, представлены на рис. 14.2, а рассчитанные в данном режиме показатели – в табл. 14.2–14.6.

- Этот пример решен также в файле Эксель «Ex 14.1», который можно скачать с моего сайта

Рис. 14.2


Регрессия
Входные данные
Входной интервал Y:
$B$3:$B$8
Входной интервал X:
$C$3:$E$8
Метки
Константа - ноль
Уровень надежности:
95
%
Параметры вывода
Выходной интервал:
$B$11
Новый рабочий лист:
Новая рабочая книга
Остатки
Остатки
Стандартизованные остатки
График остатков
График подбора
Нормальная вероятность
График нормальной вероятности
OK
Отмена
Справка

Анализ табл. 14-2

В табл. 14.2 сгенерированы результаты по регрессионной статистике. Эти результаты соответствуют следующим статистическим показателям:

- *Множественный R* – коэффициенту корреляции R;
- *R-квадрат* – коэффициенту детерминации R^2;
- *Стандартная ошибка* – остаточному стандартному отклонению

$$\sigma_0 = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-(m+1)}};$$

- *Наблюдения* – числу наблюдений n.

В табл. 14.3 сгенерированы результаты дисперсионного анализа, которые используются для проверки значимости коэффициента детерминации R^2.

Таблица 14.2

	B	C
11	ВЫВОД ИТОГОВ	
12		
13	Регрессионная статистика	
14	Множественный R	0,997
15	R-квадрат	0,995
16	Нормированный R-квадрат	0,991
17	Стандартная ошибка	5,050
18	Наблюдения	6

Таблица 14.3

	B	C	D	E	F	G
20	Дисперсионный анализ					
21		df	SS	MS	F	Значимость F
22	Регрессия	2	13962,33	6981,16	273,74	0,0004
23	Остаток	3	76,51	25,50		
24	Итого	5	14038,83			

Анализ табл.14-3

Столбцы табл. 14.3 имеют следующую интерпретацию:

1. Столбец *df* – число степеней свободы.

Для строки *Регрессия* число степеней свободы определяется количеством факторных признаков m в уравнении регрессии $k_Ф = m$.

Для строки *Остаток* число степеней свободы определяется числом наблюдений n и количеством переменных в уравнении регрессии $m + 1$: $k_О = n-(m+1)$.

Для строки *Итого* число степеней свободы определяется суммой $k_Y = k_Ф + k_О$.

2. Столбец *SS* – сумма квадратов отклонений.

Для строки *Регрессия* – это сумма квадратов отклонений теоретических данных от среднего:

$$SS_Ф^2 = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2.$$

Для строки *Остаток* – это сумма квадратов отклонений эмпирических данных от теоретических:

$$SS_О^2 = \sum_{i=1}^{n} (y_i - \hat{y})^2.$$

Анализ табл.14-3 – часть 2

Для строки *Итого* – это сумма квадратов отклонений эмпирических данных от среднего:

$$SS_Y^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad \text{или} \quad SS_Y^2 = SS_\Phi^2 + SS_O^2.$$

3. Столбец *MS* – дисперсии, рассчитываемые по формуле

$$MS = \frac{SS}{df}.$$

Для строки *Регрессия* – это факторная дисперсия σ_Φ^2.

Для строки *Остаток* – это остаточная дисперсия σ_O^2.

4. Столбец *F* – расчетное значение *F*-критерия Фишера F_p, вычисляемое по формуле

$$F_p = \frac{MS(\text{Регрессия})}{MS(\textit{Остатки})}.$$

5. Столбец *Значимость F* – значение уровня значимости, соответствующее вычисленному значению F_p. Определяется с помощью функции

$$= \text{FРАСП}(F_p; df(\text{регрессия}); df(\text{остаток})).$$

В табл. 14.4 сгенерированы значения коэффициентов регрессии a_i и их статистические оценки.

Таблица 14.4

	B	C	D	E	F	G	H	I	J
26		Коэффициенты	Стандартная ошибка	t-статистика	P-Значение	Нижние 95%	Верхние 95%	Нижние 95,0%	Верхние 95,0%
27	Y-пересечение	-1,9434	7,6254178	-0,2549	0,8153	-26,2109058	22,324	-26,21	22,324
28	Величина оборотного капитала X1, млн руб.	0,69499	0,1968595	3,5304	0,0386	0,068497396	1,3215	0,0685	1,3215
29	Стоимость основного капитала X2, млн руб.	0,20235	0,0351996	5,74857	0,0105	0,090326737	0,3144	0,0903	0,3144

Анализ табл.14-4

Столбцы табл. 14.4 имеют следующую интерпретацию:

1. *Коэффициенты* – значения коэффициентов a_i.

2. *Стандартная ошибка* – стандартные ошибки коэффициентов a_i.

3. *t-статистика* – расчетные значения t-критерия, вычисляемые по формуле

$$t\text{-статистика} = \frac{\text{Коэффициенты}}{\text{Стандартная ошибка}}.$$

5. *P-значение* – значения уровней значимости, соответствующие вычисленным значениям t_p. Определяются с помощью функции

$$=\text{СТЬЮДРАСП}(t_p;n-m-1).$$

6. *Нижние 95 %* и *Верхние 95 %* – соответственно нижние и верхние границы доверительных интервалов для коэффициентов регрессии a_i. Для нахождения границ доверительных интервалов с помощью функции = СТЬДРАСПОБР (α; $n - m - 1$) рассчитывается критическое значение t-критерия $t_{кр}$, а затем по формулам

$$\text{Нижние 95\%} = \text{Коэффициент} - \text{Стандартная ошибка} \cdot t_{кр};$$

$$\text{Верхние 95\%} = \text{Коэффициент} + \text{Стандартная ошибка} \cdot t_{кр}$$

вычисляются соответственно нижние и верхние границы доверительных интервалов.

Табл. 14.5

	B	C	D	E
33	ВЫВОД ОСТАТКА			
34				
35	Наблюдение	Предсказанная Прибыль Y, млн руб.	Остатки	Стандартные остатки
36	1	190,91	-2,91	-0,74
37	2	80,98	-2,98	-0,76
38	3	94,57	-1,57	-0,40
39	4	153,62	-1,62	-0,42
40	5	52,98	2,02	0,52
41	6	153,93	7,07	1,81

Описание табл. 14-5

В табл. 14.5 сгенерированы теоретические значения $\hat{y}_i$ результативного признака Y и значения остатков. Последние вычисляются как разность между эмпирическими y и теоретическими $\hat{y}_i$ значениями результативного признака Y.

Строим уравнение регрессии

Рассчитанные в табл. 14.4 (ячейки C27:C29) коэффициенты регрессии a_i позволяют построить уравнение, выражающее зависимость прибыли предприятий Y от величины оборотных средств X_1 и стоимости основных фондов X_2:

$$\hat{y} = -1{,}94 + 0{,}69x_1 + 0{,}20x_2.$$

Значение множественного коэффициента детерминации $R^2 = 0{,}995$ (ячейка C15 в табл. 14.2) показывает, что 99,5 % общей вариации результативного признака объясняется вариацией факторных признаков X_1 и X_2. Значит, выбранные факторы существенно влияют на прибыль предприятий, что подтверждает правильность их включения в построенную модель.

Рассчитанный уровень значимости $\alpha_p = 0{,}0004 < 0{,}05$ (показатель *Значимость F* в табл. 14.3) подтверждает значимость R^2.

Значимость коэффициента детерминации

Другой подход к проверке значимости R^2

основан на проверке попадания Fр (показатель F в табл. 14.3) в критическую область ($F^{кр}_{пр,\,\alpha}$, $+\infty$). Для рассматриваемого примера $F^{кр}_{пр,\,\alpha} = 9{,}55$, которое рассчитывается по формуле

=FРАСПОБР(0,05;C22;C23),

где в ячейке C22 вычисляется число степеней свободы $k_Ф = m = 2$, а в ячейке C23 – число степеней свободы $k_О = n-(m+1) = 6-(2+1) = 3$.

Так как $F_р = 273{,}74$ попадает в критический интервал (9,55; $+\infty$), то гипотеза $H_0 : R^2 = 0$ отвергается, т. е. коэффициент детерминации R^2 является значимым.

Показатель средней ошибки аппроксимации $\bar{\varepsilon} = 2{,}7\ \%$ также подтверждает достаточно высокую адекватность построенного уравнения. Данный показатель может быть рассчитан по формуле

{=СУММ(ABS(D36:D41)/(C3:C8))/СЧЕТ(C3:C8)*100},

где в массиве D36 : D41 табл. 14.5 рассчитаны разности между эмпирическими и теоретическими значениями результативного признака.

Значимость коэффициентов регрессии

Следующим этапом является проверка значимости коэффициентов регрессии: a_0, a_1 и a_2. Сравнивая попарно элементы массивов C27:C29 и D27:D29 (*см.* табл. 14.4), видим, что абсолютное значение свободного члена a_0 меньше, чем его стандартная ошибка. Таким образом, свободный член a_0 следует исключить из уравнения регрессии.

Стандартные ошибки коэффициентов a_1 и a_2 меньше своих стандартных ошибок. К тому же эти коэффициенты являются значимыми, о чем можно судить по значениям показателя *P-значение* в табл. 14.4, которые меньше заданного уровня значимости $\alpha = 0{,}05$.

Другой распространенный способ проверки значимости коэффициентов регрессии основан на проверке попадания $t_р$ (показатель *t-статистика* в табл. 14.4) в критическую область $(-\infty, t^{кр}_{лев,\,\alpha/2}) \cup (t^{кр}_{пр,\,\alpha/2}, +\infty)$. В генерируемых таблицах режима не приводится значение $t_{кр}$, но его можно легко вычислить с помощью функции СТЬЮДРАСПОБР. Для рассматриваемого примера значение $|t_{кр}| = 3{,}18$, которое рассчитывается по формуле

=СТЬЮДРАСПОБР(0,05;6–2–1),

где 0,05 – заданный уровень значимости;
6 – число наблюдений;
2 – число факторов в уравнении регрессии;
1 – число свободных членов в уравнении регрессии.

Значимость коэффициентов регрессии - 2

Так как $t_{\text{р}}^{a_1} = 3{,}53$ и $t_{\text{р}}^{a_2} = 5{,}75$ попадают в критический интервал $(-\infty; -3{,}18) \cup (3{,}18; +\infty)$, то коэффициенты регрессии a_1 и a_2 являются значимыми.

Подводя итог предварительному анализу уравнения регрессии, можно сделать вывод, что его целесообразно пересчитать без свободного члена a_0, который не является статистически значимым.

Для пересчета уравнения регрессии в диалоговом окне **Регрессия** необходимо задать те же самые параметры (*см.* рис. 14.2), за исключением лишь того, что следует активизировать флажок *Константа-ноль*. В случае если незначимым является коэффициент при факторном признаке, следует пересмотреть набор признаков в уравнении регрессии.

ВЫВОД ОСТАТКА

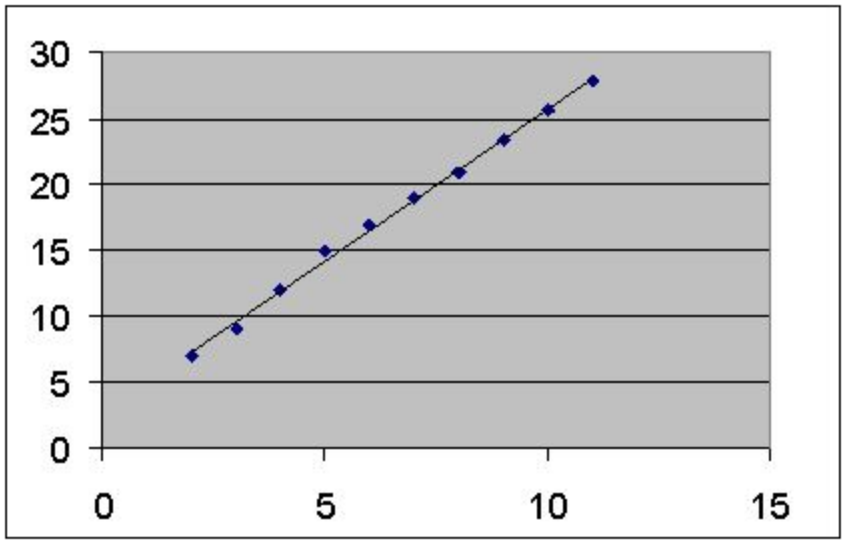
Таблица 3. Остатки

Наблюдение	Предсказанное Y	Остатки	Стандартные остатки
1	9,610909091	-0,610909091	-1,528044662
2	7,305454545	-0,305454545	-0,764022331
3	11,91636364	0,083636364	0,209196591
4	14,22181818	0,778181818	1,946437843
5	16,52727273	0,472727273	1,182415512
6	18,83272727	0,167272727	0,418393181
7	21,13818182	-0,138181818	-0,34562915
8	23,44363636	-0,043636364	-0,109146047
9	25,74909091	-0,149090909	-0,372915662
10	28,05454545	-0,254545455	-0,636685276

- При помощи этой части отчета мы можем видеть отклонения каждой точки от построенной линии регрессии. Наибольшее абсолютное значение *остатка* в нашем случае - 0,778, наименьшее - 0,043. Для лучшей интерпретации этих данных воспользуемся графиком исходных данных и построенной линией регрессии, представленными на рисунке. Как видим, линия регрессии достаточно точно "подогнана" под значения исходных данных.
- Следует учитывать, что рассматриваемый пример является достаточно простым и далеко не всегда возможно качественное построение регрессионной прямой линейного вида.

Исходные данные и линия регрессии

- Рисунок 1



- Осталась нерассмотренной задача оценки неизвестных будущих значений зависимой переменной на основании известных значений независимой переменной, т.е. задача прогнозирования.
- Имея уравнение регрессии, задача прогнозирования сводится к решению уравнения Y= x*2,305454545+2,694545455 с известными значениями x. Результаты прогнозирования зависимой переменной Y на шесть шагов вперед представлены в таблице 4.

Прогноз

Таблица 4. Результаты прогнозирования переменной Y

x	Y(прогнозируемое)
11	28,05455
12	30,36
13	32,66545
14	34,97091
15	37,27636
16	39,58182

Выводы

- Таким образом, в результате использования регрессионного анализа в пакете Microsoft Excel мы:
- построили уравнение регрессии;
- установили форму зависимости и направление связи между переменными - положительная линейная регрессия, которая выражается в равномерном росте функции;
- установили направление связи между переменными;
- оценили качество полученной регрессионной прямой;
- смогли увидеть отклонения расчетных данных от данных исходного набора;
- предсказали будущие значения зависимой переменной.
- Если *функция регрессии* определена, интерпретирована и обоснована, и оценка точности регрессионного анализа соответствует требованиям, можно считать, что построенная модель и прогнозные значения обладают достаточной надежностью.
- Прогнозные значения, полученные таким способом, являются средними значениями, которые можно ожидать.