Руководство к решению задач и контрольные задания

по дисциплине

«Корреляционный и регрессионный анализ в экономических расчетах»

для студентов бакалавров заочного обучения

по направлению «Экономика» и «Менеджмент»

Указание

1. Разделы учебного пособия, которые необходимо проработать перед выполнением контрольных заданий (стр. 3-43).
2. Выполнить контрольные задания 1-5 (стр. 43-46)
3. Для выполнения контрольных заданий 1-5 рекомендуется максимально использовать табличный процессор MS Excel, статистические функции MS Excel и пакет анализа данных MS Excel.

Отчет о лабораторной работе должен содержать разделы

1) титульный лист (см. файл-приложение);
2) описание задания (задачи);
3) таблица исходных данных;
4) описание результатов расчетов выполнения и анализ результатов (по этапам);
5) итоговое изложение полученных результатов.

Отчетность по работе

1. Отчет представляется в бумажном и электронном виде.
2. Текст отчета в электронной форме в формате MS Word отправить на mm_ugntu@mail.ru. Имя файла Фамилия_Аббревиатура группы: **Иванов_БЭГз.docx.**
3. Защита контрольной работы включает постановку целей и задачи метод расчетов и ответы на контрольные вопросы.
4. Защита проверенного отчета состоит в обосновании студентом выводов, сделанных им на основе результатов выполненных компьютерных расчетов.

2

1. Линейная регрессия и корреляция

Регрессия и корреляция широко используется при анализе связей между явлениями. Прежде всего, в экономике – исследование зависимости объемов производства от целого ряда факторов: размера основных фондов, обеспеченности предприятия квалифицированным персоналом и других; зависимости спроса или потребления населения от уровня дохода, цен на товары и т.д. Экономические показатели являются многомерными случайными величинами.

В большинстве случаев между переменными, характеризующими экономические величины, существуют зависимости, отличающиеся от функциональных. Она возникает, когда один из факторов зависит не только от другого, но и от ряда случайных условий, оказывающих влияние на один или оба фактора. В этом случае ее называют стохастической (**корреляционной)** и говорят, что переменные **коррелируют**. Виды стохастических связей между факторами могут быть линейными и нелинейными, положительными или отрицательными. Возможна такая ситуация, когда между факторами невозможно установить какую–либо зависимость.

Однако при изучении влияния одного явления на другое удобно работать именно с функциями, связывающими эти явления. Задачи построения функциональной зависимости между факторами, анализа полученных результатов и прогнозирования решаются с помощью **регрессионного анализа**.

В пособии приводятся решения задач содержащих небольшое количество данных, для того чтобы пользователь мог быстро ввести значения в таблицу Excel. Каждое решение содержит подробную инструкцию. Сначала рассмотрите пример и проверьте результаты. Затем примените пошаговые инструкции к собственному множеству данных.

1.1. КОРРЕЛЯЦИОННАЯ ЗАВИСИМОСТЬ

Для изучения зависимости между двумя числовыми переменными (x и y) сначала строят графики рассеяния. В Excel данный вид графиков называется точечной диаграммой. Используя графическое представление, можно сделать вывод о корреляционной зависимости или независимости рассматриваемых данных. Если в массиве данных присутствуют «выбросы», то их следует исключить из рассмотрения, если это возможно сделать, или усреднить, используя соседние элементы.

Теперь можно выдвинуть предположение о существовании линейной или нелинейной зависимости между переменными. Для этого найдите коэффициент корреляции и проверьте его значимость.

Тесноту линейной зависимости изучаемых явлений оценивает линейный **коэффициент парной корреляции r_{xy}:**

$$r_{xy} = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \sigma_y}, \qquad 1.1$$

где **cov(x,y)** обозначают смешенный момент второго порядка (1.5), который называется **ковариацией.**

Ковариация является мерой взаимосвязи случайных величин и может служить для определения направления их изменения:

если cov(x,y) > 0, то случайные величины изменяются в одном направлении;
если cov(x,y) < 0, то случайные величины изменяются в разных направлениях.
Очевидными свойствами ковариации являются:

- симметричность ковариации относительно случайных чисел: cov(x,y)=cov(y,x);
- $cov(x,x) = D_x$;
- если СВ X и У независимые, то cov(x,y) = 0.

Коэффициент корреляции (1.1) является величиной безразмерной. Случайные величины X и У называют **некоррелированными**, если $r_{xy} = 0$ (**отсутствует линейная зависимость между X и У**), в противном случаем можно говорить о линейной зависимости между величинами X и У, а величины называю **коррелированными.** Свойства коэффициента корреляции:

- $r_{xx} = 1$;
- $r_{xy} = r_{yx}$;
- $-1 \le r_{xy} \le 1$.

В пакете Анализ данных есть инструменты **Ковариация** и **Корреляция**, позволяющие сделать вывод о линейной зависимости случайных величин.

Пример 1.1. Для анализа зависимости объема потребления y (у.е.) хозяйств от располагаемого ежемесячного дохода x (у.е.) отобрана выборка (n=12), представленная таблицей.

i	1	2	3	4	5	6	7	8	9	10	11	12
x	107	109	110	113	120	122	123	128	136	140	145	150
y	102	105	108	110	115	117	119	125	132	130	141	144

Постройте график рассеяния и сделайте вывод о виде функциональной зависимости между объемом потребления и ежемесячным доходом в семье.

Инструкции по выполнению задания

1. Расположите данные в столбцах таблицы так, чтобы значения x были слева, а y справа (рис.1.1).
2. Выделите диапазон ячеек.
3. Щелкните мышью по кнопке Мастер диаграмм и выберите тип Точечная. Для форматирования диаграммы удобно использовать контекстное меню, которое вызывается щелчком правой кнопки мыши на форматируемом объекте.
4. Дайте название диаграмме Корреляционное поле.
5. Расположите диаграмму на листе, содержащем данные, как показано на рис.

Применим встроенную функцию **КОРРЕЛ(диапазонX; диапазонY)** для установления линейной зависимости между переменными (рис. 1.1). Найденный коэффициент корреляции 0,99 свидетельствует о сильной линейной зависимости между объёмом потребления и уровнем доходов в семье.

Проверим значимость коэффициента корреляции. Для этого сформулируем основную и альтернативную гипотезы:

Но: $r_{xy} = 0$, коэффициент незначимый;

H_1: $r_{xy} \neq 0$, коэффициент значимый.

Для проверки гипотезы воспользуемся t–критерием и уровнем значимости 5%,

$$t_{кр} = Стьюдраспобр(0{,}05;10) = 2{,}23\,;\; t_{расч} = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}} = 22{,}19.$$

Сравнивая эти значения, сделаем вывод о том, что основная гипотеза отклоняется в пользу альтернативной, т.е. коэффициент корреляции значим. По расположению точек на рис. 1.1 можно предположить, что между x и y существует линейная зависимость: $y= b_0+b_1x$.

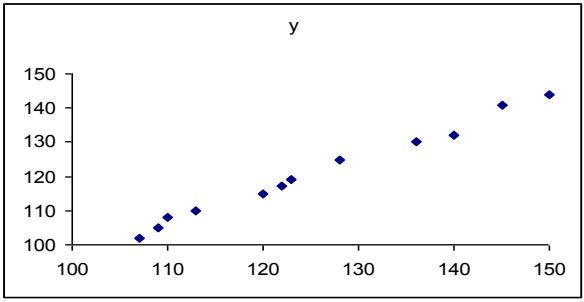


Рис 1.1. Корреляционное поле для изучения зависимости между x и y

1.2. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ ДАННЫХ

При выполнении многомерного анализа данных изучают корреляцию между каждой парой переменных. Эти результаты представляют в виде корреляционной матрицы. Инструмент анализа **Корреляция** позволяет определить парные корреляции для многих переменных. После его запуска получится нижняя треугольная часть матрицы, на диагонали которой будут стоять единицы (r_{xx}). Верхняя часть матрицы является зеркальным отражением нижней ее части, поскольку $r_{xy} = r_{yx}$.

Если надо изучить зависимость между переменными при условии управления одной или несколькими переменными, то находят коэффициенты частной корреляции. Частные коэффициенты корреляции могут оказаться полезными при определении ложных связей.

Например, изучается зависимость $y = f(x, z)$. Коэффициенты парной корреляции между x и y высокие, однако зависимость будет считаться ложной, если x линейно зависит от z. Если исключить влияние переменной z, то корреляционная зависимость между y и x может исчезнуть,

Надо найти частные коэффициенты корреляции, т.е. элиминировать один из факторов (устранить его влияние). В случае трех факторов корреляцию между y и x при элиминированном факторе z можно найти по формуле:

$$r_{yx\cdot z} = \frac{r_{yx} - r_{yz}\cdot r_{xz}}{\sqrt{1-r_{xz}^2}\cdot\sqrt{1-r_{yz}^2}}\,.$$

Подобным образом находят и остальные коэффициенты частной корреляции.

Пример 1.2.

Формируется три портфеля из десяти акций. Первый состоит из 10 акций вида А., второй содержит по 5 акций А и В; а третий включает 5 акций вида А, 3 вида В и 2 вида С. Данные о прибыли по каждому виду акций за десять месяцев представлены на рис 1.3.

Имеется ли зависимость между акциями А, В и С?

Отличаются ли данные портфели по доходности и риску?

Инструкции по выполнению задания

1. Введите данные в ячейки А1: С11, как показано на рис. 1.2.
2. В меню сервис выберите Анализ данных / инструмент Корреляция. Заполните поля диалогового окна, как показано на рис.1.3. и нажмите ОК.
3. Аналогично найдите матрицу парных ковариаций.

	A	B	C	D	E	F	G	H
1	A	B	C		Инструмент анализа Корреляция			
2	10,83	7,92	6,86			*A*	*B*	*C*
3	12,37	8,81	11,58		A	1		
4	12,79	8,18	11,58		B	0,319036	1	
5	9,43	7,14	13,23		C	0,363201	-0,05562	1
6	11,03	7,47	8,67					
7	11,45	9,41	10,65		Инструмент анализа ковариация			
8	10,59	7,11	10,99			*A*	*B*	*C*
9	8,50	8,56	8,19		A	1,591084		
10	9,60	7,96	5,55		B	0,2835	0,49629	
11	11,37	8,51	9,65		C	1,031431	-0,08822	5,0686

Рис 1.2. Данные о дивидендах по акциям за десять месяцев и Матрицы корреляции и ковариации для них

Описание результатов

Коэффициенты корреляции не очень высокие: $r_{AB} = 0,32$, $r_{AC} = 0,36$, $r_{BC} = -0,06$. Акции плохо коррелируют между собой, то есть между дивидендами по акциям существует слабая линейная зависимость.

Так как коэффициент ковариации для дивидендов по акциям В и С отрицательный, то прибыль по ним будет изменяться в разных направлениях (при увеличении дивидендов по акциям В дивиденды по акциям С будут уменьшаться). Правда, эти изменения не очень велики, около 10%.

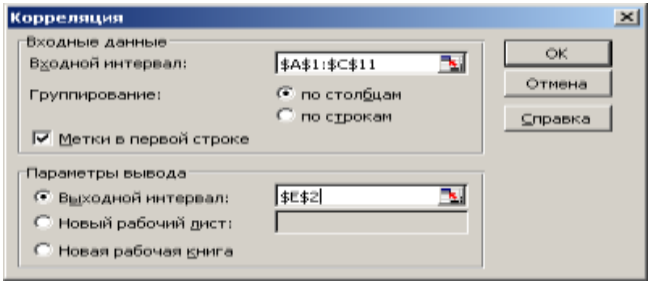


Рис 1.3. Диалоговое окно корреляции

Если рынок ценных бумаг устойчивый, то желательно исключить акции вида С из портфеля, так как cov(C, C)=5,07 наибольшая, а значит риск в их вложение высокий.

Акции А и В коррелируют слабо cov(A, B)=0,28, поэтому есть основания считать, что вложение капитала в равных долях в эти акции будет наименее рискованным. Для более правильного вывода надо вычислить дисперсии для каждого портфеля и сравнить их.

Дисперсии для первого портфеля :

$$D(10A)=100\,cov(A,A)=159.$$

Для второго: Z = 5 A + 5 B

$$D(Z) = = 66.$$

Третий портфель имеет дисперсию:

$$D(F)= D(5\,A + 3\,B + 2C)=. 25\,cov(A,A) + 9\,cov(B,B) + 4\,cov(C,C) + 30\,cov(A,B) + 20\,cov(A,C) + 12\,cov(B,C)=92,36$$

Вывод: наименьший риск получается при покупке акций А и В в равных долях.

Чтобы принять окончательное решение надо построить множество Парето, характеризующее зависимость доходности портфеля от его риска, т.е. математического ожидания и дисперсии:

M(10A)=10M(A)= 107,96; M(Z)=96,75; M(F)=98,97.

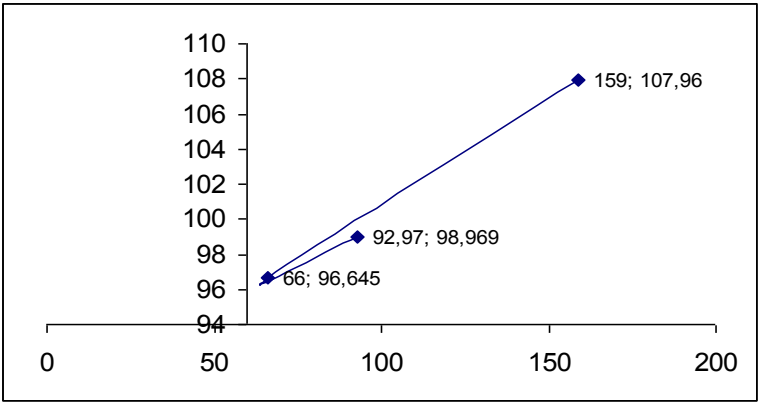


Рис. 1.4. Графическое представление доходности и риска по акциям

1.3. ПОСТРОЕНИЕ ТРЕНДА ДЛЯ ДВУХ РЯДОВ ДАННЫХ

Задача построения функциональной зависимости может быть выполнена с помощью команды **Добавить линию тренда**. В этом случае необходимо визуально исследовать зависимость между x и y и выбрать график элементарной функции, который даст лучшее приближение к экспериментальным данным. Форматирование графиков выполняется с помощью меню **Диаграмма**. Напомним, что форматируемый объект должен быть **выделен**. Существуют и другие способы форматирования: **контекстное меню** – вызывается для объекта с помощью правой клавиши мыши.

Прежде всего, надо исследовать корреляционное поле и сделать вывод о характере зависимости между переменными. Затем выполните действия (тренд построен для данных примера 1.1):

1. На диаграмме (рис.1.1) выделите маркеры, щелкнув по любой из точек данных.
2. В меню диаграмма выберите Добавить линию тренда (можно воспользоваться контекстным меню).
3. Перейдите на вкладку Тип диалогового окна Линия тренда, как показано на рис. 1.5 и выделите пиктограмму Линейный.
4. Откройте вкладку Параметры (рис.1.6) включите опции Показывать уравнение на диаграмме и Поместить на диаграмму величину достоверности аппроксимации (R^2).

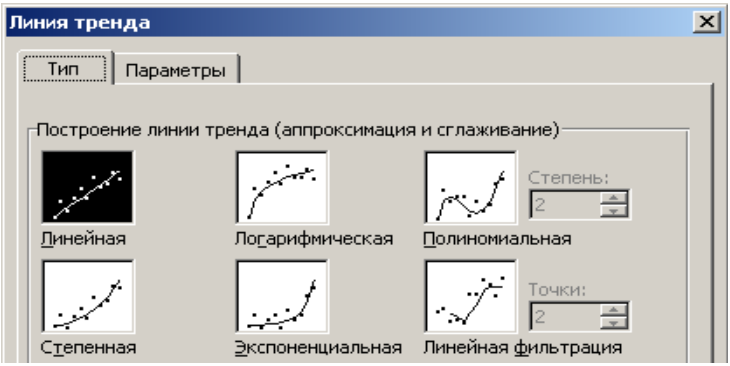


Рис. 1.5. Диалоговое окно линии тренда

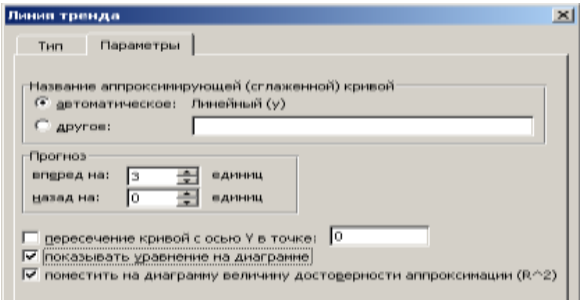


Рис 1.6. Вкладка Параметры линии тренда

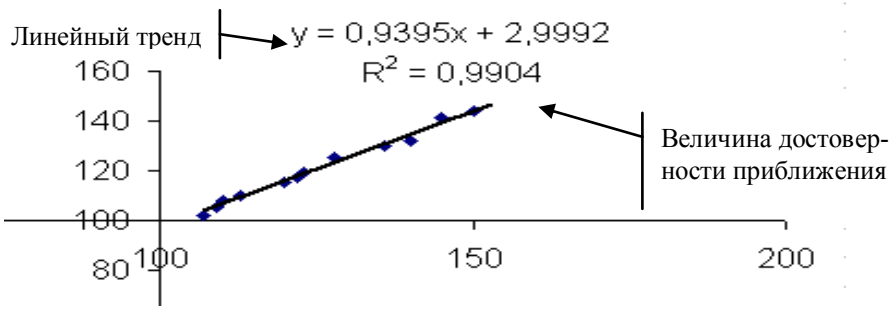


Рис. 1.7. Линейный тренд, построенный на основе данных примера 1.1

На вкладке параметры имеются и другие типы функциональных зависимостей. Предлагается самостоятельно построить остальные виды тренда и записать их уравнения. Не забывайте включать опции из пункт 4, приведенной выше инструкции.

1.4. ИНСТУМЕНТ АНАЛИЗА РЕГРЕССИЯ

Дает возможность провести более полный анализ, полученного уравнения линейного тренда с использованием методов математической статистики.

Коэффициенты уравнения линейной регрессии находятся по выборочным данным и являются величинами случайными, поэтому надо провести анализ их значимости (значимости). Надо определить значимость всего уравнения регрессии и самое главное построить прогноз по построенному уравнению, а затем провести его оценку значимости.

При построении линейного тренда предполагается, что **линейная модель** наилучшим образом характеризует зависимость между x и y:

$$y=\beta_0+\beta_1 x+\varepsilon,$$

где β_0 и β_1 **параметры** модели; ε – случайная величина (**возмущение**), характеризующая влияние неучтенных факторов.

Уравнение прямой (1.2), коэффициенты которого находят по выборочным данным, называют уравнением **регрессии** и обозначают $\hat{\mathbf{y}}$:

$$\hat{y}=b_0 + b_1\cdot x, \qquad 1.1$$

Коэффициенты регрессии b_0 и b_1 находят по методу наименьших квадратов. Они являются только **оценками** параметров модели (соответственно β_0 и β_1). Для получения **наилучших оценок** необходимо, чтобы выполнялся ряд предпосылок относительно **случайного отклонения**

$$e_i=y_i-\hat{y}_i=y_i-b_0-b_1x_i$$

индекс i означает значение факторов в одноименном испытании. Это **условия Гаусса–Маркова** (Приложение 1), а так же предположения:

- случайные отклонения имеют нормальный закон распределения;
- отсутствуют ошибки спецификации;
- число наблюдений достаточно большое: как минимум в шесть раз превышает число объясняющих факторов и другие.

Оценку b_1 называют **коэффициентом регрессии**. Ее значение показывает среднее изменение результата y с изменением фактора x на одну единицу.

Можно установить зависимость между коэффициентом регрессии и коэффициентом корреляции:

$$r_{xy} = b_1 \frac{\sigma_x}{\sigma_y} = \frac{\overline{yx}-\bar{y}\cdot\bar{x}}{\sigma_x\sigma_y} = \frac{\operatorname{cov}(x,y)}{\sigma_x\sigma_y}. \qquad 1.3$$

В качестве меры рассеивания **фактического** значения y относительно **теоретического** значения $\hat{y}$ (находится по уравнению регрессии) используется **стандартная ошибка** уравнения регрессии, которая определяется по формуле:

$$S=\sqrt{\frac{\sum e_i^2}{n-2}}. \qquad 1.4$$

ОЦЕНКА КАЧЕСТВА ПОЛУЧЕННОГО УРАВНЕНИЯ РЕГРЕССИИ СОДЕРЖИТ СЛЕДУЮЩИЕ ПУНКТЫ:

- Оценка значимости коэффициентов регрессии;
- Построение доверительных интервалов для каждого коэффициента;
- Оценка значимости всего уравнения регрессии;
- Построение прогнозного значения и доверительного интервала к ним.

Для определения **статистической значимости** коэффициентов регрессии и корреляции необходимо рассчитать **t-статистики** Стьюдента лучше всего это сделать с помощью встроенной функции СТЬДРАСПОБР [1].

ОЦЕНКА ЗНАЧИМОСТИ КОЭФФИЦИЕНТОВ РЕГРЕССИИ И КОРРЕЛЯЦИИ Устанавливает надежность полученных результатов. Случайные ошибки коэффициента корреляции и оценок параметров линейной модели вычисляются по формулам:

$$S_1 = S_{b_1} = \sqrt{\frac{S^2}{\sum_i (x_i - \bar{x})^2}} = \frac{S}{\sqrt{n}\cdot\sigma_x}$$ – **стандартное отклонение коэффициента b_1.** 1.5

$$S_0 = S_{b_0} = \sqrt{S_1^2 \cdot \frac{\sum_i x_i^2}{n}} = \sqrt{S_1^2 \cdot \overline{x^2}} = S_1 \cdot \sqrt{\overline{x^2}}$$ – **стандартное отклонение коэффициента b_0.** 1.6

$$S_r = \sqrt{\frac{1 - r_{xy}^2}{n-2}}$$.– **стандартное отклонение коэффициента корреляции.** 1.7

Любое стандартное отклонение иногда называют **стандартной ошибкой** соответствующего коэффициента.

Рассматривается основная **гипотеза о равенстве параметров регрессии нулю**.

H_0: $b_i = 0$ – коэффициент незначим;

H_1: $b_i \neq 0$ – коэффициент значимый

По выборке находят t-статистики ($T_{набл}$):

$$t_1 = T_{набл}(b_1) = \frac{b_1}{S_{b_1}}, \quad t_0 = T_{набл}(b_0) = \frac{b_0}{S_{b_0}}, \quad t_r = T_{набл}(r_{xy}) = \frac{r_{xy}}{S_r}.$$ 1.8

Критическое значение $T_{кр}$ для t-статистик находят с помощью распределения Стьюдента. Для этого надо знать **объем** выборки и задать **уровень значимости** α. Например, для α=0,05 и n=14, $\mathbf{T_{кр}=t_{\alpha/2,\,n-2}}=t_{0,025,\,12}=2{,}179$.

Выдвинутая гипотеза:

- **принимается, если выполняется неравенство** $|\mathbf{T_{набл}}| < \mathbf{T_{кр}}$ и делают вывод, что коэффициент незначим (равен нулю);
- **отвергается, если** $|\mathbf{T_{набл}}| > \mathbf{T_{кр}}$ и делают вывод, что коэффициент значим.

Более подробно о проверке гипотез можно прочитать в первой части методических указаний.

Часто при проверке качества коэффициентов используют «**грубое прав**ило»:

- если $|t| \le 1$ ($b_j < S_j$), то коэффициент статистически незначим;
- если $1 < |t| \le 2$ ($b_j < 2S_j$), то коэффициент относительно слабо значим, рекомендуется воспользоваться таблицей критических точек распределения Стьюдента;
- если $2 < |t| \le 3$, то коэффициент значим (это утверждение считается гарантированным при $\nu > 20$ и $\alpha \ge 0{,}05$);
- если $3 < |t|$, то коэффициент считается сильно значимым (вероятность ошибки при достаточном числе наблюдений не превосходит 0,001).

Каждая оценка дополняется доверительным интервалом. Для этого определяют предельную ошибку [1] для каждого коэффициента:

$$\Delta_i = t_{\alpha/2,\,n-2} \cdot S_i, \qquad 1.9$$

откуда границы доверительных интервалов находятся по формуле:

$$b_i \pm \Delta_{bi}. \qquad 1.10$$

Коэффициент **детерминации** для парной регрессии совпадает с **квадратом коэффициента корреляции $R^2 = r^2_{xy}$** и характеризует долю дисперсии **результативного** признака y, объясняемую регрессией в общей дисперсии результативного признака. Соответственно величина $1\text{-}R^2$ характеризует долю дисперсии y, вызванную влиянием неучтенных факторов в общей дисперсии признака y.

$$\underset{\text{общая сумма квадратов откл.}}{\sum (y_i - \bar{y})^2} = \underset{\text{объясненная регрессией}}{\sum (\hat{y}_i - \bar{y})^2} + \underset{\text{остаточная сумма}}{\sum (\hat{y}_i - y_i)^2} \qquad 1.11^*$$

Разделив обе части уравнения на общую сумму квадратов отклонений, получим:

$$1 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} + \frac{\sum (\hat{y}_i - y_i)^2}{\sum (y_i - \bar{y})^2},$$

$$1 = R^2 + \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2} \Rightarrow R^2 = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2}. \qquad 1.11$$

Таким образом, **коэффициент детерминации R^2** является мерой, позволяющей определить, в какой степени найденная прямая регрессии дает лучший результат для объяснения поведения зависимой переменной y, чем горизонтальная прямая $y = \bar{y}$. Очевидно, что $\mathbf{0 \le R^2 \le 1}$. Откуда следует, что чем ближе он к единице, тем больше уравнение регрессии объясняет поведение фактических значений y. Поэтому хотелось бы стремятся построить регрессию с наибольшим значением R^2.

Корень квадратный из коэффициента детерминации называется **индексом корреляции** и обозначают ρ_{xy}.

Для проверки общего качества уравнения регрессии выдвигается предположение, что коэффициенты b_0 и b_1 **одновременно равны нулю**, тогда уравнение считают незначимым, в противном случае значимым. Данная гипотеза проверяется на основе **дисперсионного** анализа, при этом сравниваются объясненная и остаточная дисперсии:

H_0: $S^2_{\hat{y}} = S^2$ – уравнение незначимо,

H_1: $S^2_{\hat{y}} > S^2$ – уравнение значимо.

Строится F-статистика:

$$F = \frac{\sum(\hat{y}_i - \bar{y})^2 / 1}{\sum(y_i - \hat{y}_i)^2 / (n-2)} = \frac{S^2_{\hat{y}}}{S^2}. \qquad 1.12$$

При выполнении условий МНК статистика имеет распределение Фишера с числом степеней свободы $\nu_1=1$, $\nu_2=n-1$. При уровне значимости α находят критичекую точку $F_{\alpha,\ 1,.n-1} = F_{кр}$ с помощью функции FИОБР и сравнивают его с наблюдаемым значением F. Так как рассматриваемая гипотеза правосторонняя [1], то:

- если $F>F_{кр}$, то гипотеза H_0 отклоняется в пользу H_1, что означает объясненная дисперсия существенно больше остаточной, следовательно, уравнение регрессии достаточно качественно отражает динамику изменения зависимой переменной от объясняющей.
- если $F<F_{кр}$, то гипотеза H_0 принимается, т.е. объясненная дисперсия соизмерима с остаточной дисперсией, вызванной случайными факторами. Это позволяет считать влияние объясняющих переменных модели несущественным, а следовательно, общее качество уравнения регрессии невысоким.

В случае **линейной регрессии** проверка нулевой гипотезы для F-статистики равносильна проверке нулевой гипотезы для t_r-статистики для коэффициента корреляции:

$$t_r = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r^2_{xy}}},$$

Можно доказать равенство:

$$F = t_r^2 = t_1^2 \qquad 1.13$$

Самостоятельную значимость коэффициент R^2 приобретает в случае **множественной** регрессии.

ПОИСК ПРОГНОЗНОГО ЗНАЧЕНИЯ И ЕГО ОЦЕНКА

Прогнозное значение $\hat{y}_p$ определяется, если в уравнение регрессии подставить значение x_p:

$$\hat{y}_p = b_0 + b_1 x_p \ . \qquad 1.14$$

Границы доверительного интервала для параметра y_p будут равны:

$$\hat{y}_p \pm t_{\alpha/2,\ n-2} \cdot S_p. \qquad 1.15$$

Чтобы найти стандартную ошибку S_p прогнозного значения $\hat{y}_p$ можно использовать два подхода: либо рассматривать параметр y_p как отдельное значение пе-

ременной x_p; или разброс y_p найти как условное среднее значение при известном значении x_p.

Доверительный интервал для отдельного значения y_p учитывает источники рассеяния: для коэффициентов регрессии (1.5, 1.6) и всего уравнения регрессии (1.4). В этом случае **стандартная ошибка прогноза S_p** вычисляется по формуле:

$$S_p = S \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}, \qquad 1.16$$

Доверительный интервал для условного среднего не учитывает дисперсию для всего уравнения регрессии (1.4), поэтому формула для вычисления ошибки прогноза имеет вид:

$$S_{\hat{y}} = \sqrt{\frac{S^2}{n} + \frac{S^2 (x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \quad . \qquad 1.17$$

Пример 1.3. Воспользуемся данными примера 1.1 для выполнения следующих заданий:

1. по данным выборок постройте линейную модель $y=\beta_0+\beta_1 x+\varepsilon$;
   a. оценить параметры уравнения регрессии $\hat{y}_x$;
   b. оценить статистическую значимость коэффициентов регрессии;
   c. оценить силу линейной зависимости между x и y;
   d. спрогнозируйте потребление при доходе $x=160$.
2. постройте модель, не содержащую свободный член $y=vx+u$.
   a. найдите коэффициент регрессии a,
   b. оценить статистическую значимость коэффициента a;
   c. оценить силу общее качество уравнения регрессии;
3. значимо или нет различаются коэффициенты b_1 и a?
4. какую модель вы выбираете?

Инструкции для выполнения примера с помощью инструмента Регрессия пакета анализ.

Для задания 1.

1. Наберите исходные данные на лист Excel, как и раньше по столбцам (рис1.1).
2. Найдите инструмент Регрессия в пакете Анализ данных и нажмите ОК. появится диалоговое окно (рис.1.8)
3. Входной интервал Y: введите ссылки на значения переменной y, включая метки диапазона.
4. Входной интервал X: введите ссылки на значения переменной x, включая метки диапазона.

5. Включите опцию Метки.
6. Включите опцию Уровень надежности и введите в поле значение 98.
7. Установите параметр вывода результатов, имя ячейки.
8. Включите опцию вывод остатков для получения теоретических значений y.
9. Нажмите ОК.
10. Появятся итоговые результаты (рис 1.9).
11. Выделите диапазон Вывод остатков и перенесите его, как показано на рис.1.9.

ВСЕ ОЦЕНКИ ПО УМОЛЧАНИЮ ПРОВОДЯТСЯ В EXCEL С УРОВНЕМ ЗНАЧИМОСТИ α=0,05 (γ=1-α=0,95)

Описание результатов по данным примера 1.1

Рисунок 1.9. состоит из четырех блоков: Регрессионная статистика, Дисперсионный анализ, данных для коэффициентов регрессии и их оценок, вывод остатков. Опишем более подробно полученные результаты.

РЕГРЕССИОННАЯ СТАТИСТИКА содержит строки, характеризующие построенное уравнение регрессии:

Для парной регрессии Множественный R равен коэффициенту корреляции (r_{xe}). По его значению 0,9952 можно сказать, что между x и y существует сильная линейная зависимость.

Строка R–квадрат равна коэффициенту корреляции в квадрате.

Нормированный R–квадрат рассчитывается с учетом степеней свободы числителя (n-2) и знаменателя (n-1) по формуле 1.11. Более подробно свойства этого коэффициента будут рассмотрены в разделе множественная линейная регрессия.

Стандартная ошибка (S) регрессии вычисляется по формуле 1.4.

Последняя строка содержит количество выборочных данных (n).

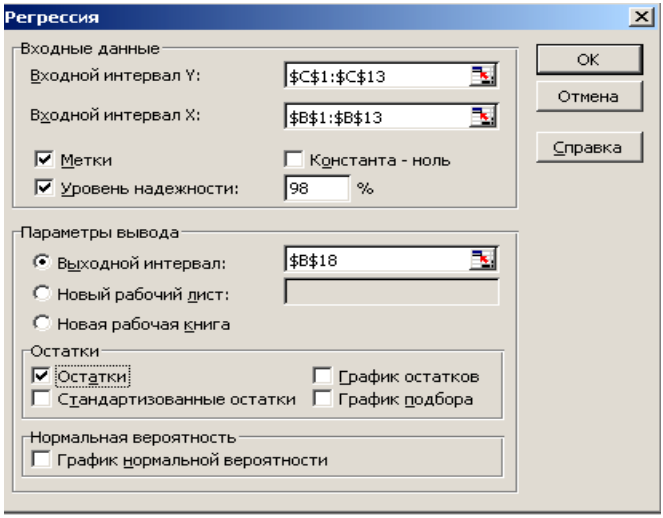


Рис 1.8. Диалоговое окно Регрессия.

Регрессионная статистика	
Множественный R	0,9952
R-квадрат	0,9904
Нормированный R-квадрат	0,9894
Стандартная ошибка	1,4228
Наблюдения	12,0000

Рис. 1.10. Фрагмент рис.1.9 Вывод итогов

ДИСПЕРСИОННЫЙ АНАЛИЗ позволяет исследовать **общую дисперсию** y (строка ИТОГО), **дисперсию для теоретических данных** (строка Регрессия) и **остаточную дисперсию** (строка Остаток).

Второй столбец (df) содержит число степеней свободы для каждой из сумм формулы 1.11*.

В третьем столбе (SS) находятся суммы квадратов (1.11*).

Четвертый столбец (MS) содержит средние значения SS/df для регрессии и остатков.

В пятом столбце вычисляется по выборочным данным значение статистика F (1.12). Последний столбец, содержит F–значение равное $P(F > F_{набл})$= FРАСП($F_{набл}$;1;10) с уровнем значимости 0,05. С его помощью можно оценить значимость всего уравнения регрессии. Это значение можно считать вероятностью выполнения гипотезы H_0. В нашем случае она практически равна нулю, следовательно, построенное уравнение дает хорошее приближение к исходным данным.

ПОСТРОЕНИЕ УРАВНЕНИЯ РЕГРЕССИИ И ОЦЕНКА ЗНАЧИМОСТИ ЕЕ КОЭФФИЦИЕНТОВ

Этот блок состоит из трех строк:

названия столбцов – первая строка

Y – пересечение – содержит все характеристики для коэффициента b_0;

третья строка (X) содержит все характеристики для коэффициента b_1.

В столбце **коэффициенты** находятся их значения b_1=0,9395 и b_0=2,9992 , используя их можно записать уравнение линейной регрессии: $\hat{y}$=2,9992+0,9395x.

Столбец **Стандартная ошибка** содержит значения S_0=3,67 и S_1=0,03 (1.6; 1.5).

В столбце **t–статистики** находятся значения, вычисленные по выборочным данным: t_0=0,81, а t_1=32,12 (1.8). По «**грубому правилу**» можно сделать вывод, что b_1 сильно значимый коэффициент, а b_0 незначим.

Подтвердить эти выводы можно с помощью данных ст**олбца Р–значение**. В этом столбе вычисляются вероятности $P(|T| > t_{набл})$ = СТЬЮДРАСП($t_{набл}$;10;2), которое можно считать вероятностью выполнения гипотезы H_0. Эта вероятность для b_1 равна нулю, что подтверждает вывод, сделанный по грубому правилу. Для коэффициента b_0 с надежностью 43% случаев можно говорить о его незначимости.

Доверительные интервалы строятся для коэффициентов по умолчанию с доверительной вероятностью 95%. Границы интервалов находятся в столбцах Нижнее 95%, Верхнее 95%:

для b_1: $0{,}8743<\beta_1<1{,}0046$;

для b_0: $-5{,}2149<\beta_0<11{,}2132$.

Так как нами была включена опция уровень надежности 98%, то получены доверительные интервалы и для этого значения $\gamma = 0{,}98$:

Нижние 98,0%	*Верхние 98,0%*
-7,18945668	13,18776625
0,85862145	1,020300894

Рис. 1.11. Доверительные интервалы с надежностью 98%

Описания, приведенные выше, практически позволили ответить на все вопросы *задания 1*, кроме построения прогнозного значения и доверительного интервала для него. Выполнить это задание можно с помощью блока вывод остатков и функции ТЕНДЕЦИЯ() или непосредственно по формулам (1.14 –1.18).

Прогнозируемое потребление при доходе $x_p = 160$ составит для данной модели:

$$\hat{y}(160) = 2{,}9992 + 0{,}9395 \cdot 160 = 153{,}31.$$

Границы доверительного интервала **условного среднего значения** y_p (1.17):

$$153{,}31 \pm 1{,}4228 \cdot \sqrt{\frac{1}{12} + \frac{(125{,}25 - 160)^2}{2108{,}67}}.$$

Таким образом, среднее потребление при доходе 160 у.е. с надежностью 95% будет находиться в интервале (152,8993; 154б4624).

Для определения границ интервала, в котором сосредоточено не менее 95% возможных объемов потребления при неограниченно большом числе наблюдений и уровне дохода $x=160$, воспользуемся формулой (1.16):

$$153{,}31 \pm 1{,}4228 \cdot \sqrt{1 + \frac{1}{12} + \frac{(125{,}25 - 160)^2}{2108{,}67}} \cdot$$

Получим границы интервала для прогнозного значения (151,4791; 155,б1409). Нетрудно заметить, что он включает в себя интервал для среднего потребления.

Коэффициент b_1 может трактоваться как предельная склонность к потреблению. Фактически он показывает, на какую величину изменится объем потребления, если предполагаемый доход возрастет на единицу.

Свободный член b_0 уравнения регрессии определяет прогнозируемое значение y при величине располагаемого дохода x, равной нулю (т.е. автономное потребление). В нашем примере $b_0=2{,}9992$ говорит о том, что при нулевом располагаемом доходе расходы на потребление составят 2,99992 у.е. Это можно объяснить для отдельных хозяйств (каждое может тратить накопленные или одолженные деньги), но для совокупности хозяйств коэффициент теряет смысл.

Следует помнить, что полученное уравнение регрессии отражает лишь общую тенденцию в поведении рассматриваемых переменных. Индивидуальные значения могут отклоняться от модельных.

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множеств	0,995188119
R-квадрат	0,990399393
Нормиров	0,989439332
Стандартн	1,422830969
Наблюден	12

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1	2088,422187	2088,422187	1031,600822	2,01529E-11
Остаток	10	20,24447966	2,024447966		
Итого	11	2108,666667			

	Коэффициенты	*Стандартная ош*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 98,0%*	*Верхние 98,0%*
Y-пересеч	2,999154781	3,686491084	0,813552701	0,434844456	-5,214859197	11,213169	-7,18945668	13,18776625
x	0,939461173	0,029249806	32,11854328	2,01529E-11	0,874288543	1,0046338	0,85862145	1,020300894

ВЫВОД ОСТАТКА

Наблюдение	*Предсказанное*	*Остатки*
1	103,5215	-1,521500264
2	105,400423	-0,40042261
3	106,339884	1,660116218
4	109,158267	0,841732699
5	115,734496	-0,73449551
6	117,613418	-0,613417855
7	118,552879	0,447120972
8	123,250185	1,749815108
9	130,765874	-0,765874274
10	134,523719	-2,523718965
11	139,221025	1,778975172
12	143,918331	0,081669308

Рис 1.9. Вывод итогов при использовании пакета Анализ данных/ Регрессия по данным примера 1.2.

Задание2.

Рассмотрим модельное уравнение, не содержащее свободного члена: $y=\beta x+\varepsilon$, тогда соответствующее ему уравнение регрессии: $\hat{y}=ax+u$.

Проведем исследование этого уравнения, так же как и в *задании* 1. Запустим инструмент Регрессия. Для заполнения полей диалогового окна (рис.1.8) повторите действия 3 – 6 из задания 1; **обязательно** включите опцию **Константа ноль** и измените параметры выходного интервала так, чтобы вывод итогов задания 1 и задания 2 не пересекались.

Вывод итогов в этом случае представлен на рис 1.12. Строка, соответствующая свободному члену уравнения, содержит **запись #Н/Д**, так как он отсутствует в уравнении.

Проведите описание результатов **самостоятельно** для полученного уравнения регрессии $\hat{y}=0{,}9631x$ также как в задании 1.

Обратите внимание, что столбцы Верхнее 95% и Нижнее 95% повторяются, так как опция уровень надежности отключена.

Задание 3.

Проверим значимо или нет, различаются коэффициенты b_1 и a. Для этого сформулируем гипотезу о равенстве математических ожиданий:

H_0: $M(b_1)=M(a)$ – коэффициенты совпадают, значимого различия нет;
H_1: $M(b_1)\neq M(b)$ – коэффициенты различаются значимо.

Для проверки гипотезы построим статистику

$$T_{\text{набл.}}=\frac{b_1-a}{\sqrt{(n-2)S_{b_1}^2+(n-1)S_a^2}}\cdot\sqrt{\frac{n^2(2n-3)}{2n}}=\frac{0{,}939-0{,}9631}{\sqrt{10\cdot 0{,}009+11\cdot 1{,}029\cdot 10^{-5}}}\sqrt{\frac{144\cdot 21}{24}}=-0{,}90112.$$

Сравним наблюдаемое значение с критическим при уровне значимости $\alpha=0{,}05$ и числом степеней свободы $\nu=2\cdot 12-2-1=21$.

Найдем критическое значение с помощью встроенной функции Стьюдента $t=2{,}080$. Поскольку $|T_{\text{набл.}}|<t$, то нет оснований для отклонения нулевой гипотезы. Это дает основания утверждать, что различия в коэффициентах незначимо.

Задание 4.

Необходимо сравнить коэффициенты детерминации двух уравнений, значения которых возьмите из отчетов Вывод Итогов (рис.1.9, рис.1.10):
для первого уравнения $R^2=0{,}9904$,
для второго уравнения $R^2=0{,}9897$.

Так как для первого уравнения это значение больше, чем для второго, то можно предположить, что первое уравнение $\hat{y}=2{,}9992+0{,}9395x$ описывает поведение зависимой переменной лучше, чем второе $\hat{y}=0{,}9631x$, так как её коэффициент детерминации больше. Сравнение двух уравнений регрессии с помощью F–статистики будет рассмотрено в разделе множественная линейная регрессия.

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,99486882
R-квадрат	0,98976396
Нормированный R-квадрат	0,89885487
Стандартная ошибка	1,40079189
Наблюдения	12

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1	2087,082	2087,082	1063,6343	1,73E-11
Остаток	11	21,5844	1,962218		
Итого	12	2108,667			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	0	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д
x	0,96310927	0,003208	300,1807	7,04E-23	0,956048	0,970171	0,956048	0,970171

ВЫВОД ОСТАТКА

Наблюдение	*Предсказанное y*	*Остатки*
1	103,0527	-1,05269
2	104,9789	0,021089
3	105,942	2,05798
4	108,8313	1,168652
5	115,5731	-0,57311
6	117,4993	-0,49933
7	118,4624	0,53756
8	123,278	1,722013
9	130,9829	-0,98286
10	134,8353	-2,8353
11	139,6508	1,349156
12	144,4664	-0,46639

Рис 1.12. Итоги расчета параметров регрессии при b_0 равном нулю. Строка У–пересечение не должна содержать значений.

1.4. МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

Как правило, на изучаемый фактор Y оказывает влияние не один, а несколько факторов X_i. Например, спрос зависит не только от цены товара, но и от доходов потребителей, а также от цены на замещающие его товары и других факторов.

Пусть зависимая переменная Y в n наблюдениях определяется m объясняющими факторами $X=(X_1, X_2 .., X_m)$, а функциональная зависимость между ними имеет вид линейной модели:

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\ldots..+\beta_mX_m+\varepsilon \qquad 1.18$$

или для индивидуальных наблюдений i,где i=1, 2, ….n

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\ldots\ldots+\beta_mx_{im}+\varepsilon_i. \qquad 1.19$$

Уравнение регрессии для индивидуальных наблюдений:

$$\hat{y}_i= b_0+b_1x_{i1}+b_2x_{i2}+\ldots\ldots+b_mx_{im}, \text{ где} \qquad 1.20$$

$\beta=(\beta_j)$, j=0,…m– вектор неизвестных параметров,

$B=(b_j)$ – вектор оценочных параметров,

$Y=(y_i)$ – i=1,…,n вектор значений зависимой переменной,

$X=(x_{ij})$ – матрица значений независимых переменных, где x_{ij} – значение переменной X_j в i–том наблюдении,

$\varepsilon=(\varepsilon_i)$ – случайные возмущения,

$E=(\mathbf{e_i= y_i – \hat{y}_i})$ случайный вектор отклонений теоретических значений $\hat{y}_i$ от фактических y_i.

Тогда уравнение (1.18) можно записать в матричном виде:

$$Y=X\cdot\beta+\varepsilon, \qquad 1.21$$

а так же уравнение (1.20):

$$\hat{Y}= XB. \qquad 1.22$$

Чтобы найти коэффициенты линейной регрессии (1.20), надо решить уравнение (1.22) относительно матрицы B. Для этого умножают обе части матричного уравнения (1.22) на транспонированную матрицу X^T. и из полученного уравнения:

$X^T Y= X^T X$ находят: $B=(X^T X)^{-1} X^T Y.$ 1.23

Полученное решение справедливо для уравнений регрессии с произвольным количеством объясняющих факторов (m), где $(X^T X)^{-1}$ обратная матрица к матрице X^TX.

Решение (1.23) уравнения регрессии (1.22) можно найти:

1. с использованием методов матричной алгебры;
2. с помощью встроенных функций Excel для работы с массивами: МОБР(), ТРАНСП(), МУМНОЖ();
3. применить инструмент анализа Регрессия.

Первый способ изучается в курсе Математика и для его реализации необходимо записать все матрицы, характеризующие уравнение 1.23.

Для реализации второго способа коэффициенты этих матриц надо занести на лист Excel, а затем применить правила работы с массивами данных.

Необходимо помнить, что матрицы для этих методов имеют вид:

$$Y = \begin{pmatrix} y_1 \\ .. \\ .. \\ y_n \end{pmatrix} \quad B = \begin{pmatrix} b_0 \\ b_1 \\ .. \\ b_m \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_{11} & .. & x_{1m} \\ 1 & x_{21} & .. & x_{2m} \\ .. & .. & .. & .. \\ 1 & x_{n1} & .. & x_{nm} \end{pmatrix}. \qquad 1.24$$

Матрица X в первом столбце содержит единицы, которые являются коэффициентом при неизвестном b_0 линейной регрессии 1.20.

Наиболее простым является последний способ поиска коэффициентов регрессии 1.20. Рассмотрим его применение на примере.

Пример 1.4. Анализируется объем сбережений Y населения за 10 лет. Предполагается, что его размер y_i в текущем году зависит от величины x_{i-1} располагаемого дохода X в предыдущем году и от величины c_i реальной процентной ставки C в рассматриваемом году. Статистические данные приведены в таблице:

Год	900	01	02	03	04	05	06	07	08	09	2010
X(тыс. руб.)	100	110	140	150	160	160	180	200	230	250	260
C, %	2	2	3	2	3	4	4	3	4	5	5
Y(тыс. руб.)	20	25	30	30	35	38	40	38	44	50	55

Задание:

1) найдите коэффициенты линейной регрессии $Y=b_0+b_1X+b_2C$;
2) оцените статистическую значимость найденных коэффициентов регрессии b_0, b_1, b_2;
3) оцените силу влияния факторов на объем сбережений населения;
4) постройте 95% -е доверительные интервалы для найденных коэффициентов;
5) вычислите коэффициент детерминации R^2 и оценить его статистическую значимость при α=0,05;
6) рассчитайте коэффициенты частной корреляции;
7) определите, какой процент разброса зависимой переменной объясняется данной регрессией;
8) найдите скорректированным коэффициент детерминации $\bar{R}^2$ и сравните его с коэффициент детерминации R^2.
9) оцените предельную склонность граждан к сбережению. Существенно ли отличается она от 0,5?
10) определите, увеличивается или уменьшается объем сбережений с ростом процентной ставки; будет ли ответ статистически обоснованным;
11) спрогнозируйте средний объем сбережений в 2011 году, если предполагаемый доход составит 270 тыс. руб., а процентная ставка будет равна 5,5%.
12) выводы по качеству построенной модели;

Все расчеты выполним с помощью ППП Excel.

Инструкции для выполнения

1. Наберите исходные данные на лист Excel, как и раньше по столбцам (рис1.13).
2. Найдите инструмент Регрессия в пакете Анализ данных и нажмите ОК, появится диалоговое окно (рис.1.8)
3. Входной интервал Y: введите ссылки на значения переменной в столбце С, включая метки диапазона.
4. Входной интервал X: введите ссылки на значения переменной в столбцах X и Y, включая метки диапазона.
5. Включите опцию Метки.
6. Включите опцию Уровень надежности и введите в поле значение 99.
7. Установите параметр вывода результатов, имя ячейки.
8. Включите опцию вывод остатков для получения теоретических значений y.
9. Нажмите ОК.
10. Появятся итоговые результаты (рис 1.14).

	A	B	C	D
1	Год	X	C	Y
2	2000	100	2	20
3	2001	110	2	25
4	2002	140	3	30
5	2003	150	2	30
6	2004	160	3	35
7	2005	160	4	38
8	2006	180	4	40
9	2007	200	3	38
10	2008	230	4	44
11	2009	250	5	50
12	2010	260	5	55

Рис.1.13. Статистические данные для анализа сбережений населения

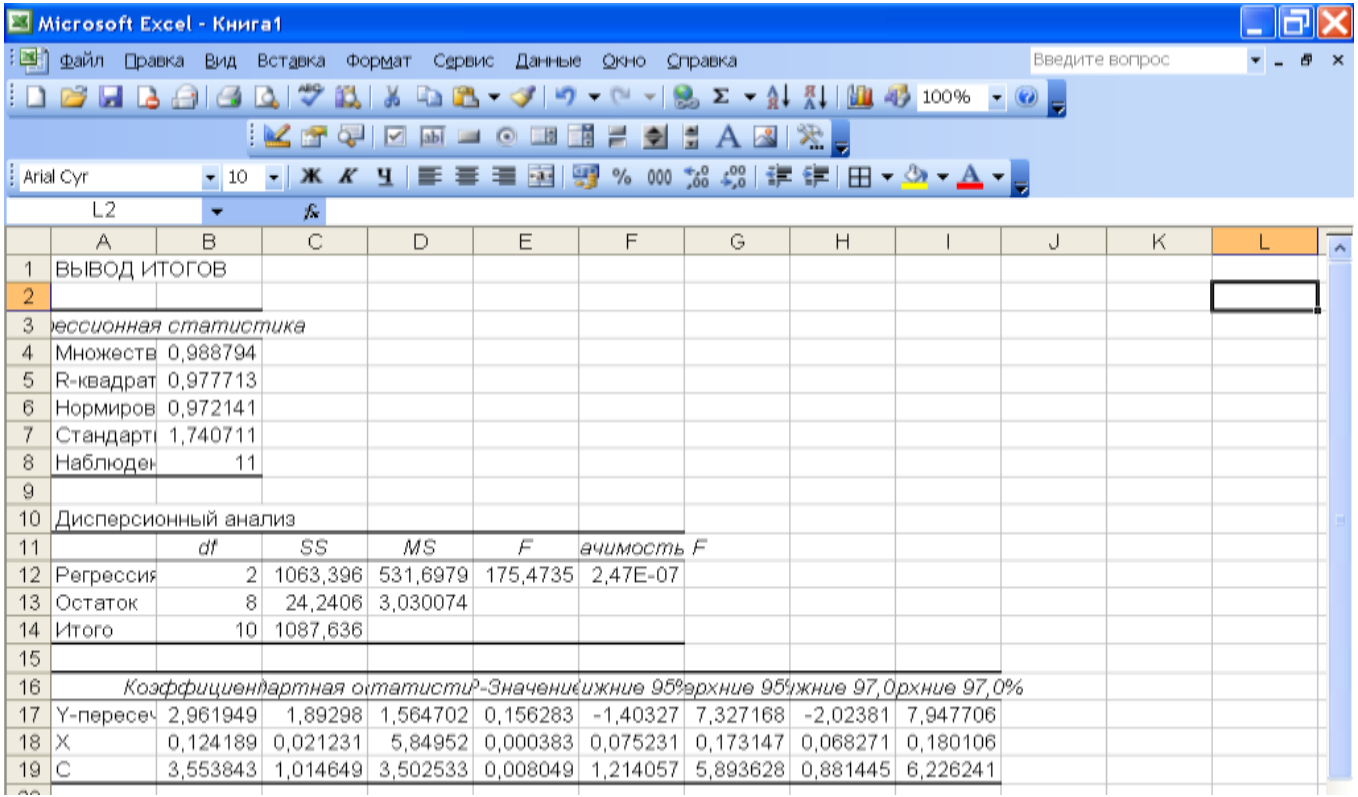
Описание результатов

УРАВНЕНИЕ ЛИНЕЙНОЙ РЕГРЕССИИ

Используя столбец Коэффициенты, запишем **уравнение регрессии**:

$$Y = 2{,}961949 + 0{,}124189 \cdot X + 3{,}553843 \cdot C.$$

При изменении доходов в предшествующем году на одну тысячу рублей сбережения увеличатся на 120 рублей, если экономическая ситуация будет стабильной. При увеличении процентной ставки на 1% сбережения могут увеличиться на 350 рублей.



	A	B	C	D	E	F	G	H	I	J	K	L
1	ВЫВОД ИТОГОВ											
2												
3	*рессионная статистика*											
4	Множеств	0,988794										
5	R-квадрат	0,977713										
6	Нормиров	0,972141										
7	Стандарті	1,740711										
8	Наблюден	11										
9												
10	Дисперсионный анализ											
11		*df*	*SS*	*MS*	*F*	*ачимость F*						
12	Регрессия	2	1063,396	531,6979	175,4735	2,47E-07						
13	Остаток	8	24,2406	3,030074								
14	Итого	10	1087,636									
15												
16	*Коэффициен*	*артная о*	*татистиі*	*-Значение*	*ижние 95%*	*ерхние 95%*	*жние 97,0*	*рхние 97,0%*				
17	Y-пересеч	2,961949	1,89298	1,564702	0,156283	-1,40327	7,327168	-2,02381	7,947706			
18	X	0,124189	0,021231	5,84952	0,000383	0,075231	0,173147	0,068271	0,180106			
19	C	3,553843	1,014649	3,502533	0,008049	1,214057	5,893628	0,881445	6,226241			

Рис. 1.14. Применение Регрессии из Анализа данных для получения параметров множественной регрессии.

ЗНАЧИМОСТЬ КОЭФФИЦИЕНТОВ РЕГРЕССИИ.

Значение t- статистик находятся в столбце с одноименным названием:

$t_0=b_0/S_0$	1,564701824
$t_1= b_1/S_1$	5,849519736
$t_2= b_2/S_2$	3,502533093

Используя «грубое правило», можно сделать вывод, что коэффициенты b_1, b_2 значимы, так как они превышают значение три. Коэффициент b_0 относительно слабо значим. Убедится в этих выводах можно используя СТЬЮДРАСПОБР(), с помощью которой найдите критические точки и постройте двухстороннюю критическую область. Для различных уровней значимости:

α=0,05 $t_{кр} = t_{0,025,\,8}$=2,306;

α=0,01 $t_{кр} = t_{0,005,\,8}$= 3,355.

Этот же вывод получите, если исследуете показания столбца P–значение. Коэффициент b_0 существенного влияния на переменную C не оказывает, т.е. может быть исключен из модели. Однако, учитывая, что в экономике, свободный член отражает экзогенную среду, лучше его оставить в уравнении регрессии, так как наличие свободного члена в линейном уравнении может только уточнить вид зависимости.

Значение t–статистики для коэффициента Y–пересечение обычно не используется.

СРАВНЕНИЕ КОЭФФИЦИЕНТОВ РЕГРЕССИИ

Простое сопоставление коэффициентов регрессии по модулю не может оценить силу влияния факторов на признак y: такое сопоставление лишено смысла. Однако их можно **нормировать (стандартизировать)**, используя формулу:

$$\alpha_j = b_j \frac{S_{xj}}{S_y},$$

где α_j- коэффициент регрессии после нормирования,

S_j – стандартная ошибка переменной x_j,

S_y – стандартная ошибка переменной y.

Нормированные коэффициенты можно сравнивать и делать вывод о влиянии факторов на переменную y. Факторы с наименьшим по модулю значением α_j оказывают на y наименьшее влияние.

Уравнение регрессии в стандартизованном масштабе имеет вид:

$$t_y = 0{,}6374\, t_x + 0{,}3817\, t_c,$$

это означает, что влияние процентной ставки (C) на объем вкладов (Y) меньше, чем влияние уровня доходов за предшествующий период (X).

ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ ДЛЯ КОЭФФИЦИЕНТОВ

Находятся в столбцах нижнее/верхнее 95%:

$-1{,}4031<\beta_0<7{,}3270$

$0{,}0753<\beta_1<0{,}1731$

$1{,}2142<\beta_2<5{,}8935$

Можно построить доверительные интервалы с уровнем надежности 97% (Рис. 1.14).

КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ

Коэффициент детерминации находится по формуле (1.11):

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2}$$

R^2 =1-24,2406/1087,636=0,9777.

Он характеризует долю разброса значений зависимой переменной Y, объясненной уравнением регрессии. В нашем примере, 98% разброса переменной Y объясняется построенным уравнением регрессии.

СКОРРЕКТИРОВАННЫЙ КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ

В случае множественной регрессии коэффициент детерминации является **неубывающей** функцией числа **объясняющих переменных**, т.е. добавление новой переменной **увеличивает значение R^2**. Поэтому при расчете коэффициента детерминации для получения несмещенных оценок в числителе и знаменателе формулы 1.11 делается поправка на число степеней свободы. Найденное значение называется **скорректированным коэффициентом детерминации:**

$$\overline{R}^2 = 1 - \frac{\sum e_i^2/(n-m-1)}{\sum (y_i - \bar{y})^2/(n-1)} \qquad 1.25$$

где

- $\sum e_i^2/(n-m-1) = S^2$ - является несмещенной **оценкой остаточной дисперсии,** т.е. дисперсией случайных отклонений точек наблюдений от линии регрессии. Ее число степеней свободы равно n-m-1, где m+1 степень свободы связана с необходимостью решения системы m+1 линейного уравнения;
- $\sum (y_i - \bar{y})^2/(n-1) = S_{\bar{y}}^2$- является несмещенной **оценкой общей дисперсии**, т.е. дисперсией отклонения Y от $\bar{Y}$, где одна степень теряется при вычислении $\bar{Y}$.

Заметим, что несмещенная оценка **объясненной дисперсии** ($S_{\hat{y}}^2$), т.е. дисперсии отклонения точек $\hat{Y}$ от $\bar{y}$, имеет m степеней свободы.

Все суммы можно найти в столбце SS дисперсионного анализа, их средние значения в столбце MS, а число степеней свободы в столбце df этого же блока.

Для нашего примера $\overline{R}^2$ =0,9721 находится в блоке регрессионная статистика в строке **нормированный**.

Можно получить формулу, устанавливающую связь между скорректированным коэффициентом детерминации и коэффициентом детерминации:

$$\overline{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-m-1}. \qquad 1.26$$

Очевидно, что:

$\overline{R}^2 < R^2$ для m >1,

$\overline{R}^2 = R^2$ только при m=1.

$\overline{R}^2$ может принимать отрицательные значения (например, если R^2=0).

Коэффициент корректируется с ростом числа объясняющих переменных.

Доказано, что **скорректированный коэффициент корреляции увеличивается при добавлении новой переменной тогда и только тогда, когда t- статистика этой переменной по модулю больше единицы.** Поэтому **добавление** в модель новых переменных осуществляется до тех пор, пока он растет.

В пакете Анализ данных приводятся значения $\overline{R}^2$ и R^2. Значимость коэффициента детерминации и скорректированного коэффициента при исследовании уравнения регрессии большая, однако, не абсолютная. При **неправильной спецификации** модели можно получить очень высокие значения этих коэффициентов, поэтому $\overline{R}^2$ и R^2 рассматриваются как один из ряда показателей, которые нужно проанализировать, чтобы уточнить строящуюся модель.

ИНДЕКС МНОЖЕСТВЕННОЙ КОРРЕЛЯЦИИ

Теснота линейной взаимосвязи в линейной регрессии выполняется с помощью индекса корреляции:

$$\rho = \sqrt{1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2}}. \qquad 1.27$$

Если X – неслучайная величина, то ρ характеризует качество подбора уравнения регрессии. Если же X – случайная переменная, то индекс корреляции является мерой тесноты линейной взаимосвязи между Y и набором факторов X.

Для нашего примера ρ=0,99 находим в строке Множественный рис 1.18.

КОЭФФИЦИЕНТЫ ЧАСТНОЙ КОРРЕЛЯЦИИ

Используются для выделения **определяющего** фактора и **второстепенных.** Необходимо определить частные зависимости между Y и X_j, при условии, что воздействие остальных факторов исключено (элиминировано). В случае трех переменных x, y, z. можно получить коэффициенты парной корреляции r_{xy}, r_{xz} и r_{yz} по формулам:

$$r_{yx\cdot z} = \frac{r_{yx} - r_{yz}\cdot r_{xz}}{\sqrt{1-r_{xz}^2}\cdot\sqrt{1-r_{yz}^2}} \quad r_{yz\cdot x} = \frac{r_{yz} - r_{yx}\cdot r_{zx}}{\sqrt{1-r_{zx}^2}\cdot\sqrt{1-r_{yx}^2}} \quad r_{zx\cdot y} = \frac{r_{zx} - r_{yz}\cdot r_{xy}}{\sqrt{1-r_{xy}^2}\cdot\sqrt{1-r_{yz}^2}}. \qquad 1.27$$

Воспользуйтесь инструкциями примера 1.2. и найдите коэффициенты парной корреляции для вычисления коэффициентов частной корреляции.

	X	*Y*	*C*
X	1		
C	0,874869	1	
Y	0,971358	0,939355	1

Рис.1.15. Коэффициенты парной корреляции для данных примера 1.3.

Анализируя, полученные данные можно сказать, что факторы X и Y дублируют друг друга (r_{xy}=0,875). Сравнивая их влияние на фактор C можно сделать вывод об исключении переменной Y из уравнения регрессии, так как $r_{xc} > r_{xc}$. Постройте уравнение регрессии, не содержащее фактор Y (C=aX). Сравните коэффициенты де-

терминации двух уравнений и сделайте вывод: следует исключать фактор Y или оставить его при построении уравнения регрессии.

ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ ПРОГНОЗА

Если уравнение регрессии имеет вид:

$$Y= b_0+b_1X_1+b_2X_2+\ldots\ldots b_mX_m,$$

то прогнозное значение вычисляется так же как в случае парной регрессии. Необходимо подставить заданные значения прогноза $X_p=(1, x_{p1}, x_{p2}, \ldots..x_{pm})$ в уравнение регрессии.

Найдем средний объем сбережений в 2011 году, если предполагаемый доход в 2010 году составит 270 тыс. рублей, а процентная ставка вырастет до 5,5%. Подставив эти значения в уравнение регрессии, получим средний объем сбережений в 2011 году: $\hat{Y}_{2011}$=56,039.

Точечная оценка объема сбережений в 2011 году может быть дополнена интервальной оценкой, полученной по формуле 1.15:

$$\hat{y}_{2011} - S_y \cdot t_{\alpha/2,n-m-1}\sqrt{X_p^T (X^T X)^{-1} X_p} < y_{2011} < \hat{y}_{2011} + S_y \cdot t_{\alpha/2,n-m-1} \cdot \sqrt{X_p^T (X^T X)^{-1} X_p}\ , 1.28$$

где $S_{\hat{y}}^2 = S^2 X^T_p (X^T X)^{-1} X_p$, $X^T_p=(1; 270; 5,5)$, $t_{0,025;\,8} = 2,306$, $S_c = 1,7407$.

Используя встроенные функции Excel, найдем матричное произведение:

$$X^T_p(X^T X)^{-1} X_p=0,457475.$$

Подставив все значения в 1.28, найдем интервальные оценки среднего сбережения населения в 2011 году: $53,324 < \hat{Y}_{2011} < 58,754$.

СКЛОННОСТЬ НАСЕЛЕНИЯ К СБЕРЕЖЕНИЮ в данной модели отражается через коэффициент b_1, определяющий на какую величину вырастет объем сбережений (b_1=0,124189) при росте располагаемого дохода на одну единицу.

Для анализа, существенно или нет коэффициент b_1=0,124189 отличается от 0,5, проверим гипотезу:

H_0: b_1=0,5

H_1: b_1<0,5

Построим t статистику, которая имеет распределение Стьюдента. Зададим уровень значимости α=0,05, число степеней свободы ν=11–2–1=8, тогда:

$T_{набл.}$=(0,124189-0,5)/0,0212=-17,7269,

$t_{0,05;\,8}$=1,860

Так как $T_{набл.} < -t_{кр}$, то H_0 должна быть отклонена. Действительно50% склонность населения к сбережениям явно завышена по сравнению с модельным значением в 12,4%.

РОСТ ПРОЦЕНТНОЙ СТАВКИ УВЕЛИЧИВАЕТ ОБЪЕМ СБЕРЕЖЕНИЙ. Эта зависимость характеризуется коэффициентом b_2= 3,553843>0. Так как коэффициент статистически значим, то ответ будет статистически обоснованным.

АНАЛИЗ КАЧЕСТВА УРАВНЕНИЯ РЕГРЕССИИ

Первое построенное по выборке уравнение редко является удовлетворительным по тем или иным характеристикам. Поэтому следующей задачей эконометрического

анализа является проверка качества уравнения регрессии. Эта проверка проводится по следующим этапам:

- проверка статистической значимости коэффициентов регрессии;
- проверка общего качества уравнения регрессии;
- проверка свойств данных: проверка выполнимости МНК.

По всем показателям нашего примера 1.3 модель может быть признана удовлетворительной:

- высокие t-статистики;
- коэффициент детерминации близок к единице;

Это означает, что модель может быть использована для целей анализа и прогнозирования. Мы не проверили выполнимость МНК и значимость коэффициента детерминации.

1.5. АНАЛИЗ ЗНАЧИМОСТИ R^2

Проверяется гипотеза об одновременном равенстве нулю всех объясняющих переменных – уравнение считается незначимым:

$$H_o: b_1= b_2=\dots b_m=0.$$

Если данная гипотеза не отклоняется, то делается вывод, что совокупное влияние всех m объясняющих переменных на зависимую переменную Y можно считать статистически незначимым, а общее качество уравнения регрессии невысоким.

Проверка данной гипотезы проводится на основе дисперсионного анализа, при этом сравниваются объясненная и остаточная дисперсии.

$$H_0: S^2_{\hat{y}} = S^2$$

$$H_1: S^2_{\hat{y}} > S^2$$

Для проверки гипотезы строится F-статистика:

$$F = \frac{\sum(\hat{y}_i - \bar{y})^2 / m}{\sum(y_i - \hat{y}_i)^2 / (n-m-1)} = \frac{S^2_{\hat{y}}}{S^2}, \qquad 1.29$$

которая при выполнении МНК имеет распределение Фишера с числом степеней свободы $\nu_1=m$, $\nu_2=n-m-1$. Критическое значение находится с помощью:

$$\text{FИОБР}() = F_{\alpha, m, n-m-1} = F_{кр} \text{ при уровне значимости } \alpha.$$

- Если $F>F_{кр}$ то гипотеза H_0 отклоняется в пользу H_1, что означает объясненная дисперсия существенно больше остаточной, следовательно, уравнение регрессии достаточно качественно отражает динамику изменения зависимой переменной от объясняющей.
- Если $F<F_{кр}$, то гипотеза H_0 принимается, т.е. объясненная дисперсия соизмерима с остаточной дисперсией, вызванной случайными факторами. Это позволяет считать влияние объясняющих переменных модели несущественным, а следовательно, общее качество уравнения регрессии невысоким.

На практике вместо указанной гипотезы проверяется, связанная с ней гипотеза о статистической значимости коэффициента детерминации R^2.

$$H_0: R^2=0$$

$$H_1: R^2>0$$

Очевидно, что если F=0, то и R^2 =0, а линия регрессии Y= $\bar{y}$ является наилучшей по МНК, т.е. величина Y` линейно не зависит от X_1, X_2, ….X_m. Анализ статистики F позволяет сделать вывод о том, что для принятия гипотезы об одновременном равенстве нулю всех коэффициентов линейной регрессии коэффициент детерминации R^2 не должен существенно отличаться от нуля. Его критическое значение уменьшается при росте числа наблюдений и может стать сколь угодно малым.

Для проверки этой гипотезы числитель и знаменатель формулы 1.29 поделим на общую сумму квадратов отклонений $\sum(y_i-\bar{y})^2$ и получим:

$$F=\frac{\sum(\hat{y}_i-\bar{y})^2/\sum(y_i-\bar{y})^2}{\sum(y_i-\hat{y}_i)^2/\sum(y_i-\bar{y})^2}\cdot\frac{n-m-1}{m}=\frac{R^2}{1-R^2}\cdot\frac{n-m-1}{m}. \qquad 1.30$$

Вернемся к результатам нашего примера 1.3. (рис.1.14).Найдем по таблице распределения Фишера критическую точку для уровня значимости α: $F_{0,05;2;8}$=4,46. Сравнивая критическое и наблюдаемое значения (F=175,47), можно сделать вывод, что коэффициент детерминации статистически значим. Это означает, что совокупное влияние переменных X и Y на переменную C существенно. Этот же вывод можно сделать по столбцу значимость F, который характеризует вероятность выполнения гипотезы H_0.

1.6. ПРОВЕРКА КАЧЕСТВА ДВУХ КОЭФФИЦИЕНТОВ ДЕТЕРМИНАЦИИ

Статистику F можно использовать и для обоснования случая **исключения** или **добавления** в уравнение регрессии k объясняющих переменных. Добавлять (исключать) переменные надо по одному.

Использовать лучше $\bar{R}^2$, так как R^2 всегда растет при добавлении новой объясняющей переменной. Зависимая переменная должна быть представлена в том же виде, что и уже существующие в исследуемом уравнении регрессии. Число наблюдений для обеих моделей должно быть одинаковым.

Пусть первоначально построенное по n наблюдениям уравнение регрессии имело вид:

$$\hat{Y}=b_0+b_1X_1+b_2X_2+.....+b_{m-k}X_{m-k}+....+b_mX_m,$$

и скорректированный коэффициент детерминации равен $\bar{R}_1^2$.

Исключим из уравнения k переменных, оказывающих наименьшее влияние на Y. По n наблюдениям построим новое уравнение регрессии:

$$Y=c_0+c_1X_1+c_2X_2+.....+c_{m-k}X_{m-k},$$

скорректированный коэффициент детерминации, для которого равен $\bar{R}_2^2$.

Необходимо определить существенно ли ухудшилось качество описания зависимой переменной Y. Для этого выдвинем гипотезы:

H_0: $\bar{R}_1^2-\bar{R}_2^2=0$ – ничего не изменилось

H_1: $\bar{R}_1^2-\bar{R}_2^2\neq 0$ – уравнение ухудшилось, если разность больше нуля.

По выборочным данным найдите статистику:

$$F = \frac{\overline{R}_1^2 - \overline{R}_2^2}{1 - \overline{R}_1^2} \cdot \frac{n - m - 1}{k}, \qquad 1.31$$

которая имеет распределения Фишера с числом степеней свободы ν_1=k, ν_2=n-m-1, где

$(\overline{R}_1^2 - \overline{R}_2^2)/k$ – потеря качества уравнения в результате того, что k переменных было отброшено. В результате появляется k дополнительных степеней свободы;

$(1 - \overline{R}_1^2)/(n - m - 1)$ – остаточная дисперсия первоначального уравнения.

Сравним критическое значение $F_{кр}$ и с наблюдаемым при уровне значимости α:

- Если $|F| > F_{кр}$, то гипотеза H_0 отклоняется в пользу H_1, что означает, одновременное исключение k объясняющих переменных существенно повлияет на качество первоначального уравнения.
- Если $|F| < F_{кр}$, то гипотеза H_0 принимается, т.е. разность $\overline{R}_1^2 - \overline{R}_2^2$ незначительная. Это позволяет считать, что исключение k объясняющих переменных модели допустимым, так как общее качество уравнения регрессии изменится несущественно.

Аналогично проверяется гипотеза о добавлении k объясняющих переменных в уравнение регрессии. В этом случае составляется статистика:

$$F = \frac{\overline{R}_2^2 - \overline{R}_1^2}{1 - \overline{R}_2^2} \cdot \frac{n - m - 1}{k}. \qquad 1.32$$

Исключим фактор C из уравнения регрессии примера 1.3. построим зависимость между Y и X. с помощью инструмента Регрессия получим уравнение:

$$\hat{Y}1 = 3{,}44 + 0{,}19\,X.$$

	J	K	L	M	N	O	P	Q	R
12	ВЫВОД ИТОГОВ			**Сравнение двух уравнений регрессии (пример 1.3)**					
13									
14	*Регрессионная статистика*			НормирR1=	0,972141				
15	Множеств	0,971358		НормирR2=	0,937262				
16	R-квадрат	0,943536		k=1					
17	Нормиров	0,937262							
18	Стандартн	2,61221		Fнабл=	8,763796				
19	Наблюден	11		Fкр=	5,317655				
20									
21	Дисперсионный анализ								
22		*df*	*SS*	*MS*	*F*	*ачимость F*			
23	Регрессия	1	1026,224	1026,22361	150,3924	6,4E-07			
24	Остаток	9	61,41275	6,82363931					
25	Итого	10	1087,636						
26									
27	*Коэффициенндартная о-статистик?-Значениижние 95%ерхние 95%ижние 95,0%рхние 95,0%*								
28	Y-пересеч	3,442259	2,833249	1,21495094	0,255296	-2,967	9,851514	-2,967	9,851514
29	X	0,189245	0,015432	12,2634586	6,4E-07	0,154336	0,224154	0,154336	0,224154

Рис. 1.16. Проверка качества двух уравнений регрессии

Коэффициенты и все остальные характеристики для этого уравнения регрессии можно посмотреть на рис 1.16. Сравним новое уравнений с уравнением полученным ранее.

$$\hat{Y} = 2{,}961949 + 0{,}124189 \cdot X + 3{,}553843 \cdot C.$$

В ячейке N18 находится значение F-статистики вычисленное по формуле 1.31. Критическое значение (ячейка N19) находится с помощью встроенной функции Excel при уровне значимости 0,05:

$$F_{кр} = \text{FРАСПОБР}(0{,}05;1;8) = 5{,}32.$$

Сравнивая эти два значения делаем вывод, что гипотеза H_0 отклоняется в пользу гипотезы H_1, то есть новое уравнение ухудшило качество приближения к выборочным данным.

1.7. ПРОВЕРКА ГИПОТЕЗЫ О СОВПАДЕНИИ УРАВНЕНИЙ РЕГРЕССИИ ДЛЯ ДВУХ ВЫБОРОК.

Необходимо сравнить два уравнения регрессии для отдельных групп наблюдений, т.е. будет одним и тем же уравнение регрессии для этих выборок. Для проверки этой гипотезы используется тест Чоу.

Пусть имеются две выборки объемом n_1 и n_2 . Для каждой из этих выборок получено уравнение регрессии:

$$\hat{Y} = b_0 + b_1 X_{1k} + b_2 X_{2k} + \ldots + b_m X_{mk} + E_k,\ k=1,2.$$

Суммы квадратов отклонений y_i от линий регрессии обозначим S_1 для первого и S_2 для второго уравнения регрессии.

Выдвинем гипотезу о равенстве соответствующих коэффициентов регрессии H_0: $b_{i1}=b_{i2}$, i=0, 1, 2, ….,m.

Объединим обе выборки в одну. Для выборки объема n_1+n_2 найдем еще одно уравнение регрессии, сумму квадратов отклонений которой обозначим S_0.

Тогда для проверки гипотезы H_0 строится статистика:

$$F = \frac{S_0 - S_1 - S_2}{S_1 + S_2} \cdot \frac{n_1 + n_2 - 2(m+1)}{m+1}, \qquad 1.33$$

которая имеет распределение Фишера с числом степеней свободы $\nu_1=m+1$, $\nu_2=n_1+n_2-2m-2$.

Если $S_0 \approx S_1+S_2$, то значение F-статистики приближается к нулю, а это значит, что уравнения регрессии обеих выборок практически одинаковые. А дальше сравним наблюдаемое и критическое значения F и делаете вывод принимается или отклоняется гипотеза H_0.

Данные исследования отвечают на вопрос, можно ли за рассматриваемый период времени построить единое уравнение регрессии или же нужно разбить его на части и для каждого временного интервала построить свое уравнение регрессии.

1.8. ПРОВЕРКА ВЫПОЛНИМОСТИ МНК. АВТОКОРРЕЛЯЦИЯ ОСТАТКОВ. СТАТИСТИКА ДАРБИНА–УОТСОНА.

Все предыдущие рассуждения основаны на том, что выполняются предпосылки МНК: мы предполагали, что случайные отклонения e_i являются независимыми случайными величинами со средней, равной нулю. При работе с фактическими данны-

ми, такое допущение не всегда выполняется. Например, если вид функции выбран неудачно, то отклонения от регрессии вряд ли будут независимыми. В этом случае замечается концентрация положительных или отрицательных отклонений от регрессии и можно сомневаться в их случайном характере.

Если последовательные значения e_i коррелируют (зависят) между собой, то говорят, что имеет место **автокорреляция остатков.**

МНК в случае автокорреляции дает несмещенные и состоятельные оценки, однако полученные в этом случае доверительные интервалы имеют мало смысла в силу своей ненадежности. Значительная автокорреляция говорит о том, что спецификация модели неправильная. Проверка остатков на автокорреляцию должна выполняться обязательно. Наиболее простым приемом обнаружения автокорреляции является метод Дарбина–Уотсона (DW). Идея, которого состоит в том, что проверяются на коррелированность не любые, а только соседние величины e_i. Соседними обычно считаются соседние по возрастанию объясняющей переменной X (в случае перекрестной выборки) или по времени (в случае временных рядов) значения e_i.

Статистика DW рассчитывается по формуле:

$$d = DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}. \qquad 1.34$$

При условии что $M(e_i)=0$, (i=1,2,….n) и n большое число можно предположить $\sum_{i=1}^{n} e_i^2 \approx \sum_{i=2}^{n} e_{i-1}^2$, тогда после преобразования получим:

$$d = \frac{2(\sum_{i=2}^{n} e_i^2 - \sum_{i=2}^{n} e_i e_{i-1})}{\sum_{i=1}^{n} e_i^2} = 2 - 2\frac{\sum_{i=2}^{n} e_i e_{i-1}}{\sum_{i=1}^{n} e_i^2}. \qquad 1.35$$

Очевидно, что $0 \le d \le 4$, так как коэффициент корреляции $\left| r_{e_i e_{i-1}} \right| \le 1$

- d=2, если $cov(e_i, e_{i-1})=0$ – автокорреляция отсутствует;
- d=0 –полная положительная автокорреляция;
- d=4 –полная отрицательная автокорреляция.

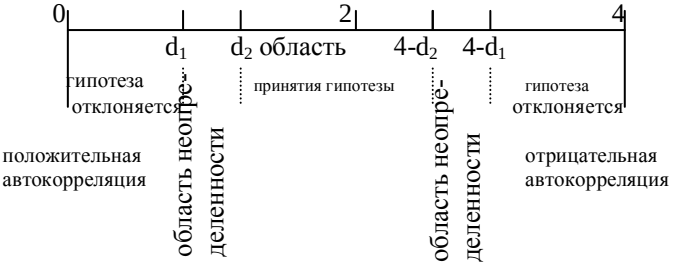


Рис. 17. Область для принятия гипотезы об отсутствии автокорреляции.

Возникает вопрос, какие значения d можно считать близкими к 2? Для обнаружения границ наблюдений статистики d существуют специальные таблицы. Для заданных α- уровня значимости; n- числа наблюдений и m –числа объясняющих переменных указывается два числа: d_1- нижняя граница и d_2- верхняя граница. Не обращаясь к таблице критических точек DW можно воспользоваться правилом, если $1,5<d<2,5$, автокорреляция отсутствует. Изобразим на рисунке числовой отрезок , используемый для проверки гипотезы об отсутствии автокорреляции.

Статистику DW для примера 1.3 находим по формуле (1.35):

$$DW=41,87394/24,2406=1,72743.$$

Для вычисления этой статистики запустите инструмент Регрессия, включив опции **Остатки** и **График остатков**, как показано на рис. 1.18. В результате получите значение случайных отклонений e_i и их графики, которые Excel строит для каждой независимой переменной, как показано на рис.1.20 и 1.21. Чтобы найти DW, можно использовать функции СУММКВРАЗН и СУММКВ.

Если зависимость между C и X линейная, то график остатков должен иметь случайный вид. На рис.1.21 видим систематический рисунок, поэтому скорее всего между C и Y существует нелинейная зависимость, а значит надо изменить модель, включая в нее нелинейную зависимость.

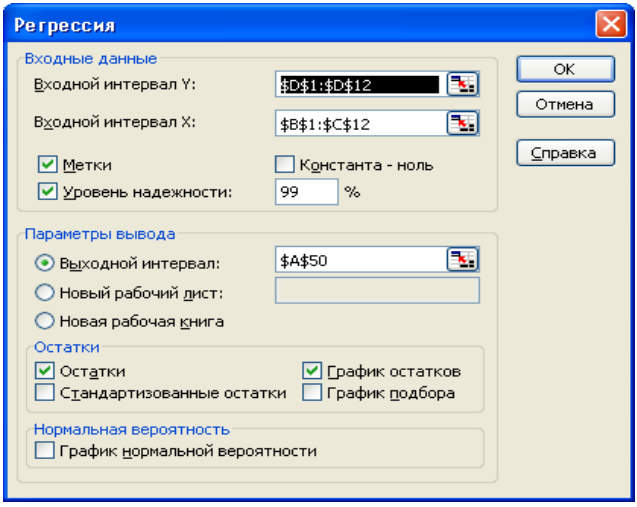


Рис. 1.18. Окно Регрессии для изучения автокорреляции остатков

Для проверки статистической значимости DW надо воспользоваться таблицей критических точек Дарбина-Уотсона, например, при уровне значимости $\alpha=0,05$ и числе наблюдений n=11 $d_1=0,658$; $d_2=1,604$. Можно считать, что автокорреляция отсутствует, так как найденная статистика попадает в критический интервал: $1,604<DW<2,396$, что является подтверждением высокого качества модели.

	A	B	C	D	E
72	ВЫВОД ОСТАТКА				
73					
74	*аблюден*	*редсказанное*	*Остатки*	ei-e(i-1)	(ei-e(i-1))2
75	1	22,48852	-2,48852		
76	2	23,73040854	1,269591	3,758111458	14,123402
77	3	31,009917	-1,00992	-2,279508462	5,1961588
78	4	28,6979627	1,302037	2,311954296	5,3451327
79	5	33,49369408	1,506306	0,204268621	0,0417257
80	6	37,04753692	0,952463	-0,553842837	0,3067419
81	7	39,531314	0,468686	-0,483777083	0,2340403
82	8	38,46124825	-0,46125	-0,929934246	0,8647777
83	9	45,74075671	-1,74076	-1,279508462	1,6371419
84	10	51,77837663	-1,77838	-0,03761992	0,0014153
85	11	53,02026517	1,979735	3,758111458	14,123402
86				Сумма	41,873938

Рис. 1.19. Вычисление наблюдаемой статистики d по формуле 1.33

Рис. 1.20. График остатков регрессии

Рис. 1.21. График остатков регрессии

1.9. МУЛЬТИКОЛЛИНЕАРНОСТЬ

Увеличение числа переменных в уравнении множественной регрессии повышает точность описания взаимосвязи, однако при этом должно выполняться условие, что Xi – **объясняющие** переменные, линейно независимые величины.

Под **мультиколлинеарностью** понимают взаимосвязь объясняющих переменных регрессии. Если между переменными Xn и Xm существует функциональная зависимость (Xn= a Xm), то говорят о **строгой** мультиколлинеарности. Чаще всего между переменными существует довольно сильная корреляционная зависимость – в этом случае мультиколлинеарность называют **нестрогой.**

При строгой мультиколлинеарности решение матричного уравнения 1.22 становится невозможным, так как матрица $X^T X$ вырожденная – её определитель равен нулю.

Если же мультиколлинеарность нестрогая, то решение матричного уравнения формально можно найти, однако все оценки мало надежны.

Чтобы обнаружить мультиколлинеарность надо найти определитель матрицы $X^T X$. Вместо этого проверяется определитель матрицы межфакторной корреляции, которую получают с помощью инструмента КОРРЕЛ.

Устранение мультиколлинеарности заключается в исключении одной из двух, находящихся во взаимосвязи переменных, либо путем пересмотра структуры уравнения регрессии. Для оценки влияния факторов на результирующий фактор Y в случае используются показатели частной корреляции (1.26). Если число переменных больше трех, то для их определения удобно пользоваться формулой:

$$r_{kp} = \frac{c_{kp}}{\sqrt{c_{kk} \cdot c_{pp}}},$$

где c_{kp} коэффициенты матрицы обратной к матрице парных коэффициентов корреляции.

1.10. ГОМОСКЕДАСТИЧНОСТЬ (постоянство дисперсии случайных отклонений)

Для применения МНК требуется, чтобы дисперсия остатков была величиной постоянной. Невыполнимость этого условия называется **гетероскедастичностью** и влечёт смещенность дисперсий оценок, так как стандартная ошибка регрессии (1.4) становится смещенной.

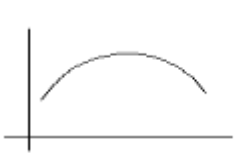
Обнаружение гетероскедастичности является сложной задачей потому что необходимо знать распределение CBY, соответствующее выбранному значению переменной x_i. В тесте Голфелда–Квандта предполагается, что стандартное отклонение пропорционально значению x_i переменной X и e_i нормально распределены, автокорреляция остатков отсутствует. Проверка на гомоскедастичность по этому тесту содержит следующие шаги:

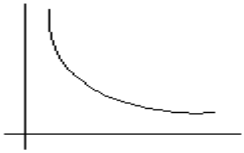
1. Все n наблюдений упорядочивают по величине.
2. Упорядоченная выборка разбивается на три подвыборки размерностью k, (n-2k) и k соответственно.
3. Центральные наблюдения исключаются из дальнейшего рассмотрения.

4. Строят регрессии для первой и последней групп и находят остаточные суммы квадратов $S1^2$ и $S2^2$ соответственно.
5. Находят их отношение $F = \frac{S2^2}{S1^2}$. Если условие гомоскедастичности выполняется, то $S1^2 = S2^2$, в противном случае $S1^2 << S2^2$.
6. Построенная F-статистика, имеет распределение Фишера с $\nu_1 = \nu_2 = k\text{-}m\text{-}1$ степенями свободы, где m число объясняющих переменных в уравнении регрессии.
7. Чем больше F превышает значение $F_{кр}$, тем более нарушена предпосылка о равенстве остаточных дисперсий.

2. НЕЛИНЕЙНАЯ РЕГРЕССИЯ

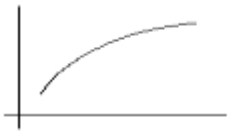
Если между экономическими явлениями существуют нелинейные соотношения, то они выражаются с помощью соответствующих функций:

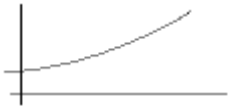




a) $y=a+bx+cx^2$

b) $y = a + \frac{b}{x}$;





c) $y=a\,x^{-b}$

d) $y=a\,b^x$

Рис.2.1. Графики некоторых нелинейных уравнений регрессии:

a) квадратичная функция (полином любой степени);
b) равносторонняя гипербола;
c) степенная;
d) показательная и др.

Кроме указанных функций для описания связи двух переменных можно использовать и другие типы кривых: $y=a+b\ln x$; $\ln y=a+bx+cx^2$ и т.д.

Различают два класса нелинейных уравнений:

1) регрессии, нелинейные относительно включенных объясняющих переменных, но линейные по оцениваемым параметрам;
2) регрессии, нелинейные по оцениваемым параметрам.

К первому классу – нелинейные по переменным – относятся кривые a и b (рис 2.1). Нелинейными по параметрам (второй класс) являются зависимости c и d на рис. 2.1.

2.1. ЛИНЕЙНЫЕ ПО ПАРАМЕТРУ

Такие модели легко приводятся к линейному виду – **линеаризуются**. Для **линейных по параметру** моделей вводят новую переменную (таблица 2.1) и переходят к построению линейной регрессии по преобразованным данным. Применяя инструмент Регрессия, к преобразованным данным можно найти все оценки параметров **преобразованных** моделей и оценить их качество.

Качество исходной модели можно оценить, используя **индекс корреляции** (1.26). Оценка статистической значимости индекса корреляции проводится с помощью F- статистики, так же как и коэффициента детерминации (1.29). Довольно часто в экономических исследованиях для оценки качества построенного уравнения используют **среднюю ошибку аппроксимации**, которая вычисляется по формуле:

$$\overline{A} = \frac{1}{n} \cdot \sum_i \left| \frac{y_i - \hat{y}_i}{y_i} \right| \cdot 100\% \qquad 2.1$$

и оценивает по модулю величину отклонений расчетных значений от фактических. **Допустимый предел значений средней ошибки аппроксимации не более 8–10%.**

Приведем примеры использования нелинейных моделей, перечисленных в таблице 2.1.

Полиномиальная модель (1) может отражать зависимость между объемом выпуска (y) и издержками производства (x); или расходами на рекламу (x) и прибылью (y) и т.д. В экономике наиболее часто используют многочлен второй степени реже третьей степени. Ограничения в применении многочленов более высоких степеней связано с требованием однородности исследуемой совокупности: чем выше степень многочлена, тем больше изгибов имеет кривая и соответственно меньше однородность по результативному признаку. Надо помнить, что графики многочленов имеют промежутки монотонности и точки экстремумов, поэтому параметры применения этих моделей не всегда могут быть логически истолкованы. Поэтому, если такая зависимость четко не определена графически (параболическая), то её лучше заменить другой нелинейной функцией.

Гиперболическая модель (2) – классическим примером этой модели является кривая Филлипса ($\beta_1>0$), характеризующая соотношение между уровнем безработицы (x) и процентом прироста заработной платы (y). При $x\to\infty$ кривая характеризуется нижней асимптотой $y=\beta_0$. Соответственно можно определить уровень безработицы, при котором заработная плата стабильна и темп её прироста равен нулю. При $\beta_1<0$ гиперболическая функция будет медленно расти для $x\to\infty$ и имеет горизонтальную асимптоту $y=\beta_0$. Такие кривые называют кривыми Энгеля, который сфор-

мулировал закономерность: с ростом доходов (x) доля доходов, расходуемых на продовольствие (y) уменьшается.

Таблица линеаризации некоторых моделей, линейных по параметру **Таблица 2.1**

№	Модель	Уравнение модели	Замена	модель
1	Полиномиальная	$Y=\beta_0+\beta_1 x+\beta_2 x^2+..+\beta_n x^n+\varepsilon$	$X=X1,\ x^2=X2,..\ ,x^n=XN$	1.18
2	Гиперболическая	$y=\beta_0+\frac{\beta_1}{x}+\varepsilon$	$X=\frac{1}{x}$	1.1
3	Полулогарифмические	$У = a + b\ \ln x + \varepsilon$ $Lny = a + b\ x + \varepsilon$	X=lnx; Y=lny	1.1

Полулогарифмические модели (3) используются, когда необходимо определить темп роста или прироста экономических показателей. Например, при анализе банковского вклада по процентной ставке, при исследовании зависимости прироста объема выпуска продукции от процентного увеличения затрат на расходы, бюджетного дефицита от темпа роста ВВП, темп роста инфляции от объема денежной массы и т.д.

2.2. НЕЛИНЕЙНЫЕ ПО ПАРАМЕТРУ

Уравнения нелинейные по параметру можно разделить на:

1) **внутренне линейные** – можно привести к линейному виду путем преобразований;
2) **внутренне нелинейные**, которые не могут быть сведены к линейной модели.

Степенная модель:

$$y=\beta_0 x^{\beta_1}\varepsilon. \qquad 2.2$$

Если прологарифмировать обе части уравнения 2.2, получится модель, легко приводящаяся к линейному виду:

$$\ln y=\ln\beta_0+\beta_1\ln x+\ln\varepsilon, \qquad 2.3$$

Надо сделать замену: Y=lny, X=lnx, A=lnβ_0, получим линейную модель (1.1).

Коэффициент модели β_1 определяет **эластичность** переменной Y по переменной X, то есть **процентное изменение Y при изменении X на 1%**. Степенная модель имеет постоянную эластичность, это легко увидеть, если продифференцировать обе части уравнения (2.3):

$$\frac{dy}{y}=\beta_1\frac{dx}{x}\Rightarrow\beta_1=\frac{dy/y}{dx/x}\Rightarrow\ \beta_1=\frac{dy}{dx}\cdot\frac{x}{y}=f'(x)\cdot\frac{x}{y} \qquad 2.4$$

Так как β_1 константа, то модель 2.3 называют моделью постоянной эластичности.

В случае парной регрессии использование обоснование использования степенной модели достаточно просто. Надо построить корреляционное поле для точек

(lnx, lny), если их расположение соответствует прямой линии, то произведенная замена хорошая и можно использовать степенную модель.

Данная модель легко обобщается на большее число переменных. Наиболее известная – производственная функция Кобба-Дугласа: $Y=\beta_0 X^{\alpha} L^{\gamma}$, где Y – объем выпуска; X – затраты капитала; L – затраты труда.

Лог-линейные модели широко используются в банковском и финансовом анализе: $Y_t = Y_0 (1+r)^t$,

где Y_0 – первоначальный банковский вклад, r – процентная ставка, Y_t – размер вклада на момент t.

Прологарифмируем обе части этой модели

$$\ln Y_t = \ln Y_0 + t \ln(1+r). \qquad 2.5$$

Введя замену $\ln Y_0=\beta_0$, $\ln(1+r)=\beta_1$, получим полулогарифмическую модель:

$$\ln Y_t = \beta_0 + \beta_1 t \qquad 2.6$$

Коэффициент β_1 в уравнении 2.6 имеет смысл **темпа прироста** переменной Y_t по переменной t, то есть характеризует относительное изменение Y_t к абсолютному изменению t. Продифференцируем 2.6 по t, получим:

$$\frac{dY_t}{Y_t} = \beta_1 dt \;\Rightarrow\; \beta_1 = \frac{dY_t / Y_t}{dt} = \frac{\text{относительное изменение } Y_t}{\text{абсолютное изменение t}}. \qquad 2.7$$

Умножив β_1 на 100%, получим темп прироста Y_t. Надо сказать, что коэффициент $\beta_1=\ln(1+r)$ определяет мгновенный темп прироста, а $r = e^{\beta_1} - 1$ характеризует темп прироста сложного процента.

Показательные модели используются, когда анализируется изменение переменной Y с постоянным темпом прироста во времени t:

$$Y = \beta_0 e^{\beta_1 t}. \qquad 2.8$$

Если провести логарифмирование, то получится уравнение аналогичное 2.5

В общем виде показательная модель имеет вид:

$$Y = \beta_0 a^{\beta_1 x}, \qquad 2.9$$

но в силу равенства $a^{\beta_1 x} = e^{\beta_1 x \ln a}$ сводится к уравнению 2.8.

2.3. КОЭФФИЦИЕНТ ЭЛАСТИЧНОСТИ

Рассматривая степенную модель, мы ввели понятие **эластичности функции**: предел отношения относительных приращений независимой переменной и зависимой называется **эластичностью** функции

$$Э = f'(x) \frac{x}{y} \qquad 2.10$$

показывает на сколько процентов изменится в среднем результат, если фактор x изменится на 1%.

Для других форм связи Э зависит от значения фактора x и не является величиной постоянной, поэтому рассчитывается **средний коэффициент эластичности**, кото-

рый показывает, на сколько процентов в среднем по совокупности изменится результат у от своей средней величины, если фактор х изменится на 1% от своего среднего значения. Формула для расчета:

$$\bar{Э} = f'(x)\frac{\bar{x}}{\bar{y}}. \qquad 2.11$$

Несмотря на широкое использование в экономике коэффициентов эластичности, возможны случаи, когда они не имеют экономического смысла. Составьте таблицу коэффициентов эластичности для всех рассмотренных нелинейных моделей самостоятельно.

2.4. ПОСТРОЕНИЕ НЕЛИНЕЙНЫХ РЕГРЕССИЙ

Можно воспользоваться командой Добавить линию тренда, так же как в случае линейного тренда (раздел 1.3): необходимо построить корреляционное поле (х, у) и выбрать одну из зависимостей на вкладке параметры: полиномиальный, логарифмический, показательный и экспоненциальный. Такой способ удобен для случая двух переменных.

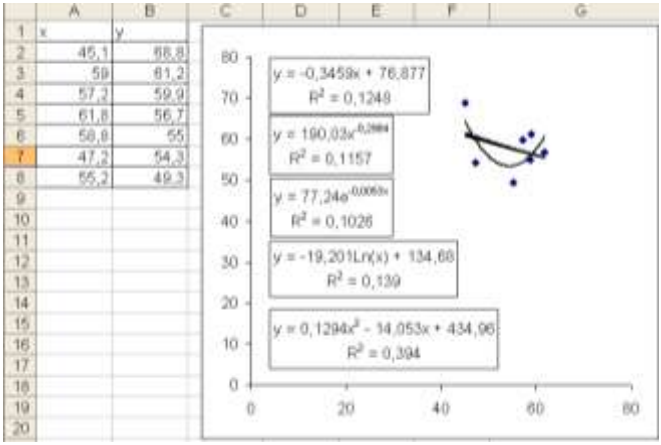
Использовать инструмент Регрессия можно только для преобразованных данных. Этот способ дает много не нужной информации.

Пример 3.1. По семи территориям Южного федерального округа за 2001 год известны значения двух признаков:

Район	Расходы на покупку продовольственных товаров в общих расходах, %, у	Среднедневная плата одного работающего, руб., х
Ростовская обл.	68,8	45,1
Краснодарский край	61,2	59,0
Ставропольская обл.	59,9	57,2
Волгоградская обл.	56,7	61,8
Ингушская респ.	55,0	58,8
Кабардино-Балкария	54,3	47,2
№ ская	49,3	55,2

Задание

1. Постройте уравнения регрессии для модели:
   a) линейной;
   b) степенной;
   c) экспоненциальной;
   d) логарифмической; гиперболы.
2. Оцените каждую модель через среднюю ошибку аппроксимации $\bar{A}$ и F-критерий Фишера.



	x	y
2	45,1	68,8
3	59	61,2
4	57,3	59,9
5	61,8	56,7
6	58,8	55
7	47,2	54,3
8	55,2	49,3

Рис 2.2. Использование линий тренда для построения нелинейных моделей.

Проще всего построить поле корреляции, а затем добавить линии тренда (см. параграф 1.3.). Для полученнх уравнений надо найти коэффициент аппроксимации и проверить F–критерий.

1a. **Уравнение линейной регрессии**: $b_0 = 76,88$ $b_1 = -0,35$

$\hat{y}=76,88-0,35\cdot x.$ $R^2 = 0,12/$

Вариация результата на 12% объясняется вариацией фактора x. F – статистику найдем по формуле1.13 **F = 0,71**

Так как $F_{набл.}=0,71 < F_{0,05;\ 1;\ 5} = 6,61$, то параметры линейного уравнения и показатель тесноты связи между x и y статистически незначимы и гипотеза о линейности уравнения регрессии отклоняется. Самостоятельно вычислите величину средней ошибки аппроксимации:

$$\overline{A} = \frac{100\%}{n}\sum\left|\frac{y-\hat{y}}{y}\right| = 8\%.$$

1.b. **Степенная модель $y=\alpha x^{\beta}$**.

$$\hat{Y} = 190,03\cdot x^{-0,298}.$$

Подставляя в уравнение регрессии фактические значения x , получим $\hat{y}_x$. По этим значениям, используя формулу для индекса корреляции (1.26), получим

$$\rho_{xy}=0,3758,$$

и среднюю ошибку аппроксимации:

$$\overline{A}=0,562761\cdot 100/7=8,04\%.$$

Характеристики степенной модели указывают, что она не намного лучше линейной функции описывает связь между x и y.

1c. Аналогично 1.b. для показательной модели $y=\alpha\cdot\beta^x$ сначала нужно выполнить линеаризацию

$$\ln y=\ln\alpha+x\cdot\ln\beta$$

и после замены переменных $Y=\ln y$, $b_0=\ln\alpha$, $b_1=\ln\beta$ рассмотрим линейное уравнение:

$$Y=b_0+b_1x.$$

Используя столбцы для Y и x из предыдущей таблицы, получим коэффициенты:

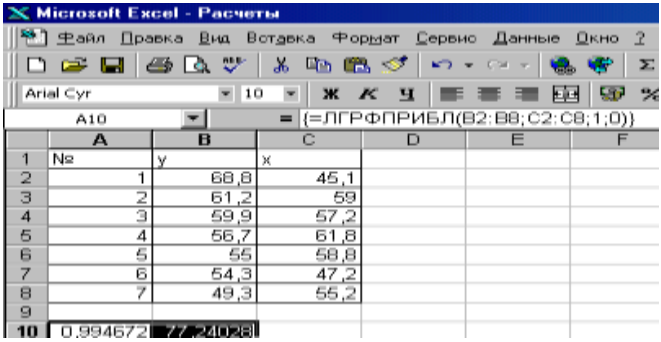
$$b_0=4{,}346921,\ b_1=-0{,}0054$$

и уравнение $Y=4{,}346921-0{,}0054\cdot x$.

После потенциирования запишем уравнение в обычной форме:

$$\hat{y}=77{,}24\cdot0{,}9947^x.$$

Все эти расчеты можно не делать, если воспользоваться для вычисления параметров b_0 и b_1 модели $y=\alpha\beta^x$ встроенной статистической функцией ЛГРФПРИБЛ. Выполните самостоятельно и сравните результаты. Убедитесь, что значения вычисленные по формулам и полученные с помощью функции ЛГРФПРИБЛ() совпадают (рис.2.4)



	A	B	C	D	E	F
1	№	y	x			
2	1	68,8	45,1			
3	2	61,2	59			
4	3	59,9	57,2			
5	4	56,7	61,8			
6	5	55	58,8			
7	6	54,3	47,2			
8	7	49,3	55,2			
9						
10	0,994672	77,24028				

Рис. 2.4. Коэффициенты показательной модели, полученные с помощью функции ЛГРФПРИБЛ

Тесноту связи оценим с помощью индекса корреляции $\rho_{xy}= 0{,}3589$, который вычисляется по формуле (1.26). Связь между x и y небольшая. Коэффициент аппроксимации, вычисленный по формуле (3.3) $\bar{A}=8\%$ говорит о повышенной ошибке приближения, но в допустимых пределах. Сравнивая, показатели степенной и показательной функций можно сделать вывод, что степенная функция чуть лучше описывает связь между x и y чем показательная.

1.d. Аналогичные расчеты надо провести и для равносторонней гиперболы $y=\alpha+\beta/x$, которая линеаризуется заменой $X=1/x$.

Для этого уравнения в таблицу исходных значений надо добавить столбец $X=1/x$, а все остальные вычисления проведите, используя один из описанных выше способов:

$b_0= 38{,}5$; $b_1= 1051{,}4$; $\hat{y}= 38{,}5+1051{,}4\cdot1/x$; $\rho_{xy}=0{,}3944$; $A=8{,}1\%$.

Получена наибольшая оценка тесноты связи по сравнению с линейной, степенной и показательной регрессиями, а $\bar{A}$ остается в пределах допустимого значения, это означает, что для описания зависимости расходов на покупку продовольственных товаров в общих расходах (у в %) от среднедневной заработной платы одного работающего (х в руб.) необходимо из предложенных моделей выбрать гиперболическую.

2. Введем гипотезу H_0: уравнение регрессии статистически незначимо и рассмотрим статистику (1.30):

$$F_{\text{набл.}} = \frac{\rho^2_{xy}}{1-\rho^2_{xy}} \cdot \frac{7-1-1}{1} = \frac{0{,}1555}{0{,}8445} \cdot 5 = 0{,}92 .$$

$F_{кр}=6{,}61$ при уровне значимости $\alpha=0{,}05$ смотри в пункте 1.а.

Гипотеза H_0 о статистической незначимости параметров уравнения принимается. Результат можно объяснить небольшим числом наблюдений и сравнительно невысокой теснотой гиперболической зависимости между х и у .

Контрольная работа для заочников.

N– последние две цифры номера зачетки;

Подставьте N в каждую задачу (1–5) и получите исходные данные своего варианта.

Задача 1 . (Смотрите решение примера 1.3).

По территориям Центрального района известны данные за 2002 г. (табл. 3).

Таблица 3.

Район	Средний размер назначенных ежемесячных пенсий, руб., у	Прожиточный минимум в среднем на одного пенсионера в месяц, руб., х
Брянская обл.	2400+N	1780+10%N
Владимирская обл.	2260+N	2020+10%N
Ивановская обл.	2210+N	1970+10%N
Калужская обл.	2260+N	2010+10%N
Костромская обл.	2200+N	1890+10%N
г. Москва	2500+N	3020+10%N
Московская обл.	2370+N	2150+10%N
Орловская обл.	2320+N	1660+10%N
Рязанская обл.	2150+N	1990+10%N
Смоленская обл.	2200+N	1800+10%N
Тверская обл.	2220+N	1810+10%N
Тульская обл.	2310+N	1860+10%N

Ярославская обл.	2290+N	2500+10%N

Задание

1. Найдите параметры уравнений линейной регрессии.
2. Оцените тесноту связи с помощью показателей корреляции.
3. С помощью F-критерия Фишера оцените статистическую надежность результатов регрессионного моделирования.
4. Рассчитайте прогнозное значение результата, если прогнозное значение фактора увеличиться на 10% от его среднего уровня. Определите доверительный интервал прогноза для уровня значимости $\alpha = 0{,}05$.
5. Выводы оформите в аналитической записке.

Задача 2. *(Смотрите решение примера 3.1).*

По территориям Центрального района известны данные за 2003 г. (табл.4.).

Таблица 4.

Район	Для денежных доходов, направленных на прирост сбережений во вкладах, займах, сертификатах и на покупку валюты, в общей сумме среднедушевого денежного дохода, %, *y*	Среднемесячная начисленная заработная плата, руб., *x*
Брянская обл.	6,9+0,01*N	2890+N
Владимирская обл.	8,7+0,01*N	3340+N
Ивановская обл.	6,4+0,01*N	3000+N
Калужская обл.	8,4+0,01*N	3430+N
Костромская обл.	6,1+0,01*N	3560+N
Орловская обл.	9,4+0,01*N	2890+N
Рязанская обл.	11,0+0,01*N	3410+N
Смоленская обл.	6,4+0,01*N	3270+N
Тверская обл.	9,3+0,01*N	3570+N
Тульская обл.	8,2+0,01*N	3520+N
Ярославская обл.	8,6+0,01*N	3810+N

Задание

1. Постройте поле корреляции и сформулируйте гипотезу о форме связи.
2. Рассчитайте параметры уравнений линейной, степенной, экспоненциальной, логарифмической, обратной, гиперболической регрессий.
3. Дайте с помощью среднего (общего) коэффициента эластичности сравнительную оценку силы связи фактора с результатом.
4. Оцените с помощью средней ошибки аппроксимации качество уравнений.
5. Оцените с помощью F-критерия Фишера статистическую надежность результатов регрессивного моделирования. По значениям характеристик, рассчитанных в п. 4 и данном пункте, выберите лучшее уравнение регрессии и дайте его обоснование.
6. Рассчитайте прогнозное значение результата, если прогнозное значение фактора увеличится на 10% его среднего уровня. Определите доверительный интервал прогноза для уровня значимости $\alpha = 0{,}05$.
7. Оцените полученные результаты, выводы оформите в аналитической записке.

Задача 3. *(Смотрите решение примера 3.1).*

Зависимость среднемесячной производительности труда от возраста рабочих характеризуется моделью: $y = a + bx + cx^2$. Ее использование привело к результатам, представленным в табл. 5

Т а б л и ц а 5.

№ п/п	Производительность труда рабочих, руб., *y*	
	Фактическая	Расчетная
1	1200+N*10	1000+N*10
2	800+N*10	1000+N*10
3	1300+N*10	1300+N*10
4	1500+N*10	1400+N*10
5	1600+N*10	1500+N*10
6	1100+N*10	1200+N*10
7	1200+N*10	1300+N*10
8	900+N*10	1000+N*10
9	1100+N*10	1000+N*10
10	900+N*10	900+N*10

Задание. Оцените:

1) качество модели, определив ошибку аппроксимации,
2) индекс корреляции,
3) F-критерий Фишера.

Задача 4. *(Решите, используя формулы 1.8 – 1.13).*

Изучается зависимость потребления материалов *y* от объема производства продукции *x*. По 20 наблюдениям были получены следующие варианты уравнения регрессии:

1. $y = 3 + \underset{(6,48)}{2}\ x + \varepsilon$,
2. $\ln y = 2,5 + \underset{(6,19)}{0,2 \cdot \ln x} + \varepsilon$, $r^2 = 0,68+0,01*n$.
3. $\ln Y = 1,1 + \underset{(6,2)}{0,8 \cdot \ln X} + \varepsilon$, $r^2 = 0,69\ 0,01*n$.
4. $Y = 3 + \underset{(3,0)}{1,5 \cdot X} + \underset{(2,65)}{0,1 \cdot X^2}$, $r^2 = 0,70,01*n$

В скобках указаны фактические значения *t*-критерия.

Задание

1. Определите коэффициент детерминации для 1-го уравнения.
2. Запишите функции, характеризующие зависимость *y* от *x* во 2-м и 3-м уравнениях.
3. Определите коэффициенты эластичности для каждого из уравнений.
4. Выберите наилучший вариант уравнения регрессии.

Задача 5. *(Смотрите пример 1.4).*

Имеются данные о деятельности крупнейших компаний некоторого государства в 2009 году (табл.6).

Таблица 6.

№	Чистый доход млрд. руб., y	Оборот капитала млрд. руб., x_1	Использованный капитал млрд. руб., x_2	Численность служащих тыс. чел., x_3
1	6,6+0,01*N	6,9*(1+N/100)	83,6+0,1*N	222,0+N
2	3,0+0,01*N	18,0*(1+N/100)	6,5+0,1*N	32,0+N
3	6,5+0,01*N	107,9*(1+N/100)	50,4+0,1*N	82,0+N
4	3,3+0,01*N	16,7*(1+N/100)	15,4+0,1*N	45,2+N
5	0,1+0,01*N	79,6*(1+N/100)	29,6+0,1*N	299,3+N
6	3,6+0,01*N	16.2*(1+N/100)	13,3+0,1*N	41,6+N
7	1,5+0,01*N	5,9*(1+N/100)	5,9+0,1*N	17,8+N
8	5,5+0,01*N	53,1*(1+N/100)	27,1+0,1*N	151+N
9	2,4+0,01*N	18,8*(1+N/100)	11,2+0,1*N	82,3+N
10	3+0,01*N	35,3*(1+N/100)	16,4+0,1*N	103+N
11	4,2+0,01*N	71,9*(1+N/100)	32,5+0,1*N	225,4+N
12	2,7+0,01*N	93,6*(1+N/100)	25,4+0,1*N	675+N
13	1,6+0,01*N	10*(1+N/100)	6,4+0,1*N	43,8+N
14	2,4+0,01*N	31,5*(1+N/100)	12,5+0,1*N	102,3+N
15	3,3+0,01*N	36,7*(1+N/100)	14,2+0,1*N	105+N
16	1,8+0,01*N	13,8*(1+N/100)	6,5+0,1*N	49,1+N
17	2,4+0,01*N	64,8*(1+N/100)	22,7+0,1*N	50,4+N
18	1,6+0,01*N	30,4*(1+N/100)	15,8+0,1*N	480+N
19	1,4+0,01*N	12,1*(1+N/100)	9,3+0,1*N	71+N
20	0,9+0,01*N	31,3*(1+N/100)	18,9+0,1*N	43+N

Задание

1. Рассчитайте параметры линейного уравнения множественной регрессии с полным перечнем факторов.
2. Оцените статистическую значимость параметров регрессионной модели с помощью t–критерия; нулевую гипотезу о значимости уравнений и показателей тесноты связи проверьте с помощью F-критерия.
3. Рассчитайте матрицы парных и частных корреляций и на их основе и по t-критерию для коэффициентов регрессии отберите существенные факторы в модели. Постройте модель только с существенными факторами и оцените её параметры.
4. Рассчитайте прогнозное значение результата, если прогнозное значение факторов составляет 80% от их максимальных значений.
5. Рассчитайте ошибки и доверительные интервал прогноза для уровня значимости 5 или 10%.
6. Оцените полученные результаты, выводы оформите в аналитической записке.

Вопросы к зачету:

1. Выборочный метод и статистическое оценивание.
2. Взаимосвязь случайных величин.
3. Свойства выборочных оценок.
4. Интервальные оценки.
5. Парная регрессия. Спецификация модели.
6. Метод наименьших квадратов.
7. Линейная регрессия и корреляция. Смысл и оценка параметров.
8. Оценка значимости параметров регрессии.
9. Оценка общего качества уравнения регрессии.
10. Множественная регрессия. Смысл и оценка параметров.
11. Коэффициент множественной корреляции. Частная корреляция.
12. Оценка надежности результатов множественной регрессии.
13. Нелинейная регрессия. Различные формы моделей.

Рекомендуемая литература:

1) И.И.Елисеева. Эконометрика. Москва. Финансы и статистика. - 2001;

1

Приложение 1.

Условия Гаусса-Маркова:

1. Математическое ожидание случайных отклонений равно нулю: $M(\varepsilon_i)=0$ для всех наблюдений.
2. Дисперсия случайных отклонений постоянная $D(e_i)=\sigma^2$.

$D(e_i)= M(e_i- M(e_i))^2= M(e^2_i) - M^2(e_i)=\sigma^2$

Если свойство постоянства дисперсии выполняется, то говорят о *гомоскедастичности* случайной величины, в противном случае СВ *гетероскедастична*.

3. Случайные отклонения e_i и e_j независимы

$$\operatorname{cov}(\varepsilon_i, \varepsilon_j)=\begin{cases} 0, \text{если } i \neq j \\ \sigma^2, \text{если } i = j \end{cases}.$$

Если условие выполняется, то говорят об отсутствии *автокорреляции*.

4. Случайные отклонения не зависимы от объясняющей переменной, то есть $\operatorname{cov}(e_i, x_i)=0$

   Для экономических моделей не столь важно.
5. Модель является линейной относительно параметров.

Теорема Гаусса-Маркова.

Если условия 1-5 выполняются, то оценки, полученные по МНК являются:

1) *Несмещенными*: $M(b_0)=\beta_0$, $M(b_1)=\beta_1$;
2) *Состоятельными*: $D(b_0)\to 0$, $D(b_1)\to 0$ при $n\to\infty$.
3) *Эффективными*: параметры b_0 и b_1 имеют наименьшую дисперсию по сравнению с другими для данных параметров.

Такие оценки называются (*BLUE*).наилучшими линейными несмещенными.

! Если условия 2 и 3 нарушаются, то свойства несмещенности и состоятельности сохраняются, а свойство эффективности нет.