

Лабораторная работа по теме «Парная регрессия и корреляция»

Формальная постановка задачи

Парная регрессия представляет собой модель однофакторной регрессии, которая аналитически выражает связь между двумя переменными – y и x , т. е. модель вида:

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

Модель позволяет находить оценки значений результативного признака с помощью функции, которая отражает корреляционную связь между переменными x и y . Фактическое значение результативного признака y в каждом отдельном случае практически складывается из двух слагаемых:

$$y = \hat{y}_x + \varepsilon,$$

здесь: $\hat{y}_x$ – теоретическое (расчётное) значение результативного признака; ε – случайная величина, характеризующая отклонение реального значения от теоретического. Чем ближе теоретические значения $\hat{y}_x$ подходят к фактическим данным y , тем меньше отклонения (остатки) ε .

Случайная величина ε обусловлена ошибками спецификации, ошибками выборки и ошибками измерения. К ошибкам спецификации относятся неверный выбор математической функции для $\hat{y}_x$ и недоучет в уравнении регрессии какого-либо существенного фактора. Ошибки выборки имеют место в силу неоднородности исходных статистических данных, что, как правило, бывает при изучении экономических процессов. Если совокупность неоднородна, то уравнение регрессии не имеет практического смысла. Для получения приемлемого результата обычно исключают из совокупности наблюдения с аномальными значениями исследуемых признаков.

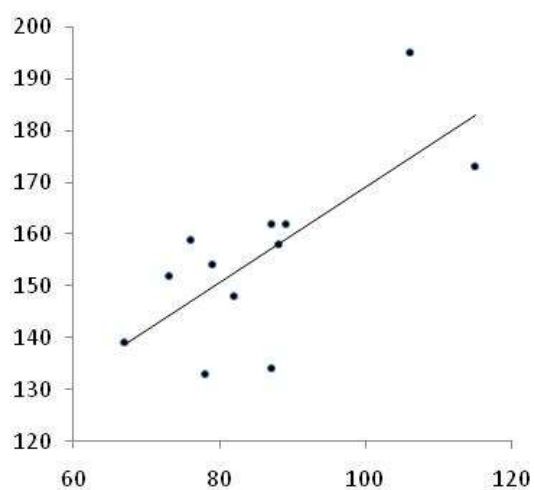
Ошибки спецификации можно уменьшить, изменяя форму модели (вид уравнения регрессии), а ошибки выборки – увеличивая объем исходных данных. Ошибки же измерения при использовании методов регрессионного анализа практически неустранимы, поскольку часто невозможно проверить достоверность получаемой статистической информации.

Предполагая, что ошибки измерения сведены к минимуму, основное внимание в эконометрических исследованиях уделяют ошибкам спецификации модели.

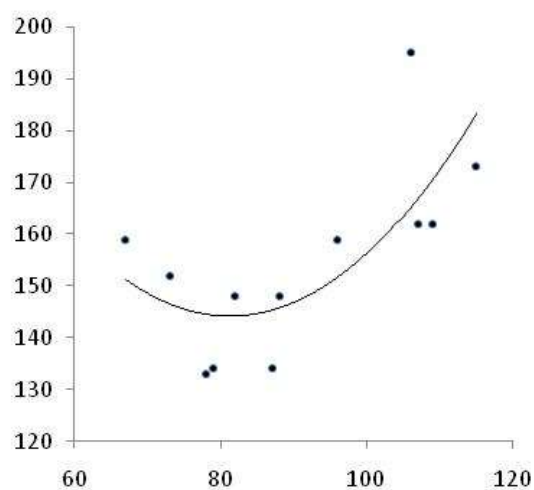
Подбор функции $\hat{y}_x = f(x)$ может быть осуществлен тремя методами:

- 1) графическим;
- 2) аналитическим;
- 3) экспериментальным.

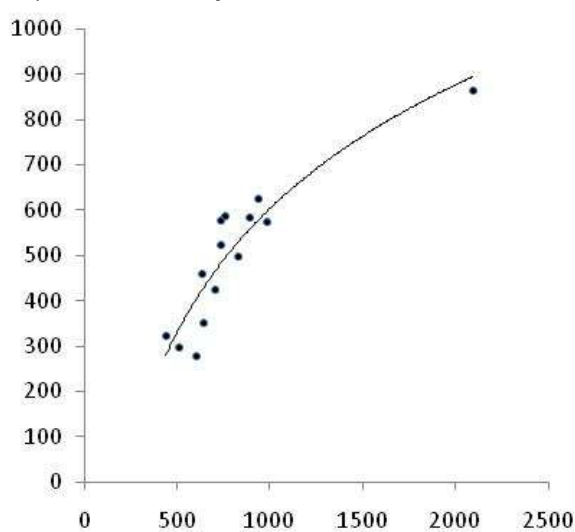
При изучении зависимости между двумя признаками наиболее простым является графический метод подбора уравнения регрессии. Он основан на поле корреляции. Основные типы кривых, используемые при количественной оценке связи двух переменных, представлены на рис. 1:



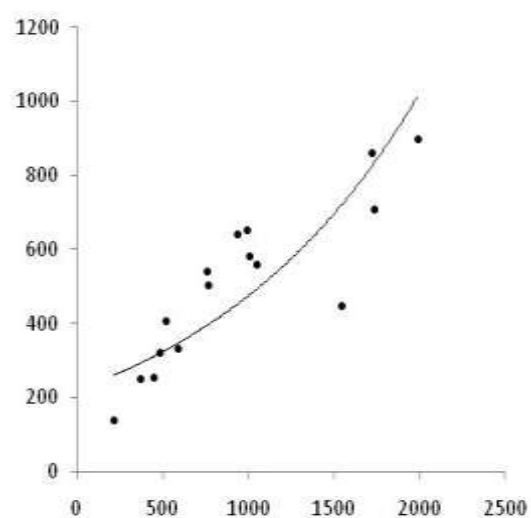
а) линейная $\hat{y} = a + bx$



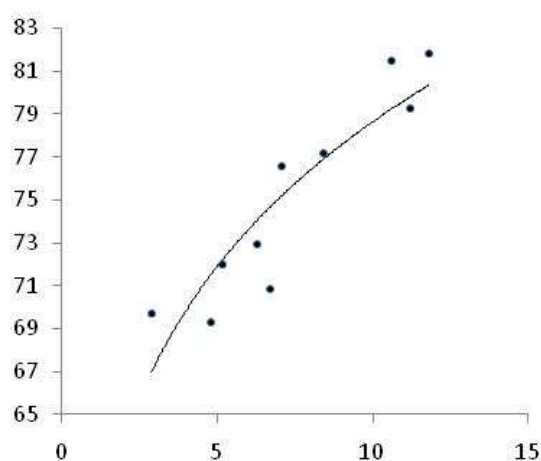
б) параболическая $\hat{y} = a + bx + cx^2$



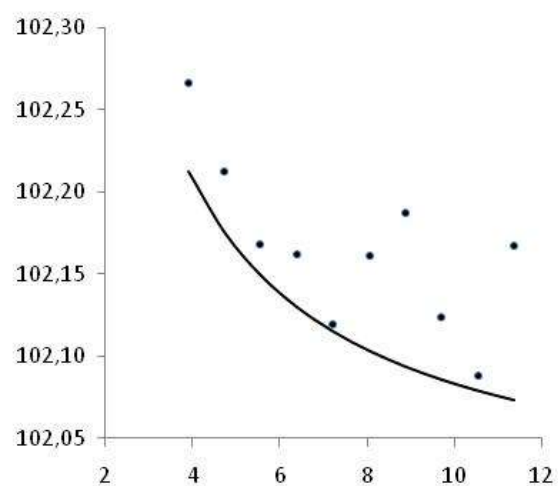
в) полулогарифмическая $\hat{y} = a + b \cdot \ln x$



г) показательная $\hat{y} = ab^x$



д) степенная $\hat{y} = ax^b$



е) гиперболическая $\hat{y} = a + \frac{b}{x}$

Рис. 1. Основные виды моделей парной регрессии

Линейная регрессия чаще других используется в эконометрических исследованиях из-за простоты расчёта параметров и возможности четкой экономической интерпретации коэффициента регрессии.

Параболическая регрессия применяется для описания процессов с монотонным развитием и отсутствием пределов роста. На практике такая зависимость может иметь место в течение некоторого непродолжительного периода.

Гиперболическую регрессию применяют для изучения зависимости удельных расходов сырья, материалов, топлива от объема выпускаемой продукции, времени обращения товаров от величины товарооборота, процента прироста заработной платы от уровня безработицы (кривая Филипса), расходов на непродовольственные товары от доходов или общей суммы расходов (кривые Энгеля).

Полулогарифмическая и показательная модели регрессии применяются при изучении процессов, которые имеют предел роста результативного показателя, например, в демографии.

Степенная регрессия также используется довольно часто, т.к. кривые спроса и предложения, производственные функции (функция Кобба-Дугласа) являются степенными функциями.

При обработке информации на компьютере выбор вида уравнения регрессии обычно осуществляется экспериментальным методом, т. е. путем сравнения величины остаточной дисперсии $\sigma_{\text{ост}}^2$, рассчитанной для разных моделей.

В практических исследованиях, как правило, имеют место отклонения фактических данных от теоретических $\varepsilon = y - \hat{y}_x$. Они обусловлены влиянием не учитываемых в уравнении регрессии факторов. Величина этих отклонений и лежит в основе расчета остаточной дисперсии:

$$\sigma_{\text{ост}}^2 = \frac{1}{n} \sum (y - \hat{y}_x)^2 = \frac{1}{n} \sum \varepsilon^2.$$

Чем меньше величина остаточной дисперсии, тем меньше влияние не учитываемых в уравнении регрессии факторов и тем лучше уравнение регрессии подходит к исходным данным. При функциональной связи линия регрессии проходит через все точки корреляционного поля, фактические значения результативного признака совпадают с теоретическими $y = \hat{y}_x$. В этом случае остаточная дисперсия $\sigma_{\text{ост}}^2 = 0$.

Считается, что число наблюдений должно в 7-8 раз превышать число рассчитываемых параметров при переменной x . Это означает, что искать линейную регрессию, имея менее 7 наблюдений, вообще не имеет смысла. Если вид функции усложняется, то требуется увеличение объема наблюдений, т.к. каждый параметр при x должен рассчитываться хотя бы по 7 наблюдениям. Значит, если мы выбираем параболу $\hat{y} = a + bx + cx^2$, то требуется объем информации уже не менее 14 наблюдений.

Классический подход к оцениванию параметров линейной регрессии $\hat{y} = a + bx$ основан на методе наименьших квадратов (МНК). МНК позволяет получить такие оценки параметров a и b , при которых сумма квадратов откло-

нений фактических значений результативного признака y от теоретических $\hat{y}_x$ минимальна:

$$\sum_{i=1}^n (y_i - \hat{y}_{x_i})^2 = \sum_{i=1}^n \varepsilon_i^2 \rightarrow \min.$$

Для линейных и нелинейных уравнений, приводимых к линейным, решается следующая система относительно a и b :

$$\begin{cases} an + b \sum_{i=1}^n x_i = \sum_{i=1}^n y_i, \\ a \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i. \end{cases}$$

Можно воспользоваться готовыми формулами, которые вытекают из этой системы:

$$b = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2}, a = \bar{y} - b\bar{x}.$$

Построению степенной модели $\hat{y} = ax^b$ предшествует процедура линеаризации, которая производится путем логарифмирования обеих частей уравнения:

$$\ln y = \ln a + b \ln x.$$

Обозначив $Y = \ln y$, $X = \ln x$, $A = \ln a$, получим линейное уравнение

$$Y = A + bX,$$

которое можно решить по аналогичным формулам.

Показательную модель $\hat{y} = ab^x$ также приводят к линейному виду путем логарифмирования

$$\ln y = \ln a + x \ln b.$$

Обозначив $Y = \ln y$, $B = \ln b$, $A = \ln a$, получим линейное уравнение

$$Y = A + Bx.$$

Уравнение гиперболы $\hat{y} = a + \frac{b}{x}$ линеаризуется заменой $z = \frac{1}{x}$. В результате также получается линейное уравнение, к которому применим МНК,

$$y = a + bz.$$

В парной линейной регрессии оценивается значимость не только уравнения в целом, но и отдельных его параметров. С этой целью для каждого из параметров определяется его стандартная ошибка: m_b и m_a .

$$m_b = \frac{S_{\text{ост}}}{\sigma_x \cdot \sqrt{n}}, \quad m_a = S_{\text{ост}} \cdot \frac{\sqrt{\sum x^2}}{\sigma_x \cdot n},$$

где $S_{\text{ост}}^2 = \frac{\sum (y - \hat{y}_x)^2}{n-2}$ – остаточная дисперсия на одну степень свободы. Затем вычисляются фактические значения t -критерия:

$$t_b = \frac{b}{m_b}, \quad t_a = \frac{a}{m_a}.$$

Их величина сравнивается с табличным значением t -критерия Стьюдента при определенном уровне значимости α и числе степеней свободы $(n - 2)$. Если фактическое значение t -критерия превышает табличное, то соответствующий параметр считается существенным.

Доверительные интервалы для параметров уравнения регрессии определяются как $a \pm t_{\text{табл}} \cdot m_a$, $b \pm t_{\text{табл}} \cdot m_b$. Поскольку знак коэффициента регрессии указывает на рост результативного признака y при увеличении фактора x ($b > 0$), уменьшение результативного признака y при увеличении фактора x ($b < 0$), или его независимость от фактора x ($b = 0$), то границы доверительного интервала для коэффициента регрессии не должны иметь разные знаки, например, $-5 \leq b \leq 7$. Это говорит о том, что истинное значение коэффициента регрессии одновременно содержит положительные и отрицательные величины и даже ноль, чего не может быть.

Чтобы иметь общее суждение о качестве модели из относительных отклонений по каждому наблюдению, определяют среднюю ошибку аппроксимации:

$$\bar{A} = \frac{1}{n} \sum A_i = \frac{1}{n} \sum \left| \frac{y - \hat{y}_x}{y} \right| \cdot 100\%.$$

Допустимый предел значений $\bar{A}$ - не более 8 - 10%.

Средний коэффициент эластичности $\bar{\varepsilon}$ показывает, на сколько процентов в среднем по совокупности изменится результат y от своей средней величины при изменении фактора x на 1% от своего среднего значения:

$$\bar{\varepsilon} = f'(\bar{x}) \cdot \frac{\bar{x}}{\bar{y}}.$$

Приведем формулы для расчета средних коэффициентов эластичности для наиболее часто используемых типов уравнений регрессии (табл. 1):

Таблица 1

Вид функции	Уравнение регрессии	Средний коэффициент эластичности, $\bar{\varepsilon}$
линейная	$\hat{y} = a + bx$	$\frac{b \cdot \bar{x}}{a + b \cdot \bar{x}}$
степенная	$\hat{y} = ax^b$	b
показательная	$\hat{y} = ab^x$	$\bar{x} \cdot \ln b$
полулогарифмическая	$\hat{y} = a + b \cdot \ln x$	$\frac{b}{a + b \cdot \ln \bar{x}}$
гиперболическая	$\hat{y} = a + \frac{b}{x}$	$-\frac{b}{a \cdot \bar{x} + b}$
параболическая	$\hat{y} = a + bx + cx^2$	$\frac{(b + 2c \cdot \bar{x}) \cdot \bar{x}}{a + b \cdot \bar{x} + c \cdot \bar{x}^2}$

Уравнение регрессии всегда дополняется показателем тесноты связи изучаемых факторов. Для линейной регрессии в качестве такого показателя выступает линейный коэффициент парной корреляции r_{xy} ($-1 \leq r_{xy} \leq 1$):

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

где σ_x, σ_y - средние квадратические отклонения величин x и y определяются по формулам

$$\sigma_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}, \sigma_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n}}.$$

Знак r_{xy} характеризует направление, а абсолютная величина r_{xy} - тесноту связи.

Качественно оценить тесноту линейной корреляционной связи между x и y можно с помощью таблицы Чеддока (табл. 2):

Таблица 2

Диапазон изменения $ r_{xy} $	Характер тесноты связи
0,1 – 0,3	слабая
0,3 – 0,5	умеренная
0,5 – 0,7	заметная
0,7 – 0,9	высокая
0,9 – 0,99	весьма высокая

Чем ближе значение $|r_{xy}|$ к единице, тем сильнее линейная связь между факторами. При значении $|r_{xy}|$, достаточно близком к 0, линейная корреляционная связь между x и y отсутствует, но при иной спецификации модели связь между признаками может оказаться достаточно тесной.

Для нелинейной регрессии в качестве показателя тесноты связи используют индекс корреляции ρ_{xy} ($0 \leq \rho_{xy} \leq 1$):

$$\rho_{xy} = \sqrt{1 - \frac{\sum (y - \hat{y}_x)^2}{\sum (y - \bar{y})^2}}.$$

Оценку качества подбора функции даёт коэффициент детерминации r_{xy}^2 для линейной регрессии и индекс детерминации ρ_{xy}^2 для нелинейной. Они характеризуют долю дисперсии результативного признака y , объясняемую регрессией, в общей дисперсии результативного признака.

Индекс детерминации ρ_{xy}^2 можно сравнивать с коэффициентом детерминации r_{xy}^2 для обоснования возможности применения линейной функции. Чем больше кривизна линии регрессии, тем величина r_{xy}^2 меньше ρ_{xy}^2 . А близость этих показателей указывает на то, что нет необходимости усложнять форму уравнения регрессии и можно использовать линейную функцию.

Индекс детерминации используется для проверки существенности в целом уравнения регрессии по F -критерию Фишера:

$$F_{\text{набл}} = \frac{\rho_{xy}^2}{1 - \rho_{xy}^2} \cdot \frac{n - m - 1}{m},$$

где ρ_{xy}^2 – индекс детерминации, n – число наблюдений, m – число параметров при переменной x . Наблюдаемое значение F -критерия сравнивается с табличным $F_{\text{табл}} = F_{\alpha; k_1; k_2}$ при уровне значимости α и числе степеней свободы $k_2 = n - m - 1$ (для остаточной суммы квадратов) и $k_1 = m$ (для факторной суммы квадратов).

Если $F_{\text{табл}} < F_{\text{набл}}$, то уравнение регрессии признается статистически значимым и надежным. Если $F_{\text{табл}} > F_{\text{набл}}$, то уравнение регрессии признается статистически незначимым.

После подтверждения существенности уравнения регрессии его можно использовать для прогнозирования. Путем подстановки в уравнение соответствующего значения факторной переменной $x = x_p$ определяется точечное прогнозное значение $\hat{y}_p$. Однако, согласно теории вероятностей, событие, состоящее в том, что непрерывная случайная величина X приняла конкретное значение x , является возможным, но обладает нулевой вероятностью. Поэтому точечный прогноз необходимо дополнить интервальной оценкой.

Рассчитывается стандартная ошибка прогноза

$$m_{\hat{y}_p} = S_{\text{ост}} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{n \cdot \sigma_x^2}},$$

и границы соответствующего доверительного интервала

$$\hat{y}_{p_{\min}} = \hat{y}_p - t_{\text{табл}} \cdot m_{\hat{y}_p}, \quad \hat{y}_{p_{\max}} = \hat{y}_p + t_{\text{табл}} \cdot m_{\hat{y}_p}.$$

Решение типовой задачи регрессионного анализа с помощью Excel

В настоящее время существует значительное количество специализированных программ для ПК, позволяющих проводить эконометрический анализ. К наиболее распространённым относятся SAS, SPSS, Stata, EvIEWS, Statgraphics и Statistica. Однако большинство этих программ достаточно сложны для освоения. Табличный процессор Excel на сегодняшний день является одним из наиболее доступных средств для решения эконометрических задач.

Excel позволяет решать эконометрические задачи двумя способами: с помощью встроенных функций, и с помощью инструмента «Анализ данных».

Рассмотрим решение типовой задачи регрессионного анализа, покажем стандартные средства Excel, которые можно при этом использовать.

Условия задачи

Приводятся данные по 12 регионам некоторой территории за год.

Таблица 3

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного (ден.ед.), x	Среднедневная заработная плата (ден.ед.), y
1	78	133
2	82	148
3	87	134
4	79	154
5	89	162
6	106	195
7	67	139
8	88	158
9	73	152
10	87	162
11	76	159
12	115	173

Задание

1. Постройте поле корреляции (диаграмму рассеяния) и сформулируйте гипотезу о форме связи исследуемых величин.
2. Для характеристики зависимости y от x рассчитайте параметры следующих функций:
 - а) линейной $\hat{y} = a + bx$;
 - б) степенной $\hat{y} = ax^b$;
 - в) показательной $\hat{y} = ab^x$;
 - г) полулогарифмической $\hat{y} = a + b \cdot \ln x$;
 - д) гиперболической $\hat{y} = a + \frac{b}{x}$.
3. Оцените статистическую значимость параметров регрессии и корреляции с помощью F -критерия Фишера и t -критерия Стьюдента
4. Выполните прогноз среднедневной заработной платы y при прогнозном значении среднедушевого прожиточного минимума x , составляющем 107% от среднего уровня. Оцените точность прогноза, рассчитав ошибку прогноза и его доверительный интервал.
5. Оцените тесноту связи с помощью индексов корреляции ρ_{xy} и детерминации ρ_{xy}^2 .
6. Дайте с помощью среднего коэффициента эластичности $\bar{\epsilon}$ сравнительную оценку силы связи фактора с результатом.
7. Оцените качество уравнений регрессии с помощью средней ошибки аппроксимации $\bar{A}$.
8. Оцените с помощью F -критерия Фишера статистическую надежность результатов регрессионного моделирования.
9. Результаты расчетов представьте в табличной форме (табл.4).

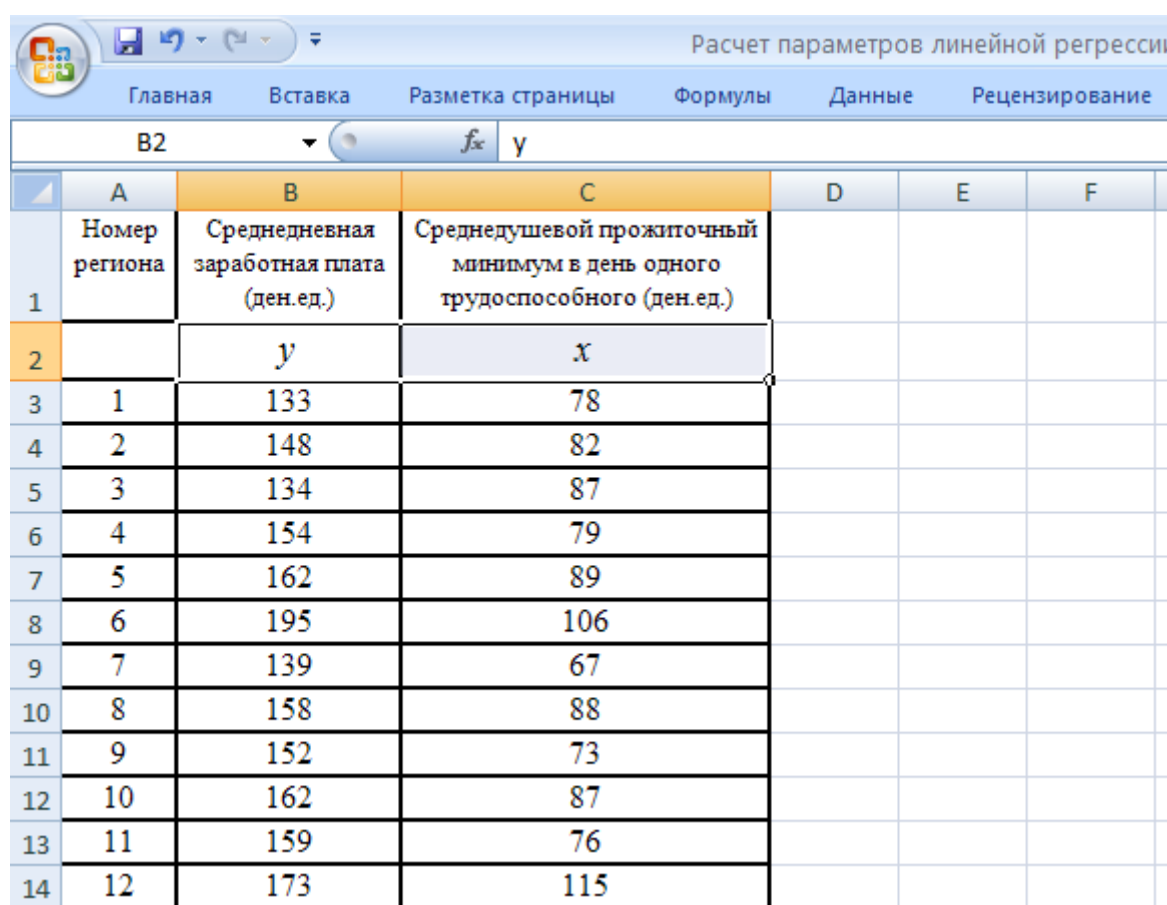
Таблица 4

Вид функции	Уравнение регрессии	Средняя ошибка $\bar{A}$	Индекс корреляции ρ_{xy}	F-критерий Фишера
линейная				
степенная				
показательная				
полулогарифмическая				
гиперболическая				

10. Сравните построенные модели по значениям рассчитанных характеристик, выберите лучшее уравнение регрессии и дайте его обоснование. Проанализируйте полученные результаты, сформулируйте выявленные закономерности.

Решение задачи

Для решения задачи линейной регрессии с помощью табличного процессора Excel создайте новую книгу с именем «Расчет параметров линейной регрессии» и на первом листе введите, начиная с ячейки A1, исходные данные в виде таблицы, как показано на рис.2. Причём для удобства дальнейших расчётов рекомендуется обозначения исследуемых величин x и y ввести в отдельную строку под строкой с наименованиями, а столбец со значениями переменной y поместить перед столбцом со значениями переменной x .

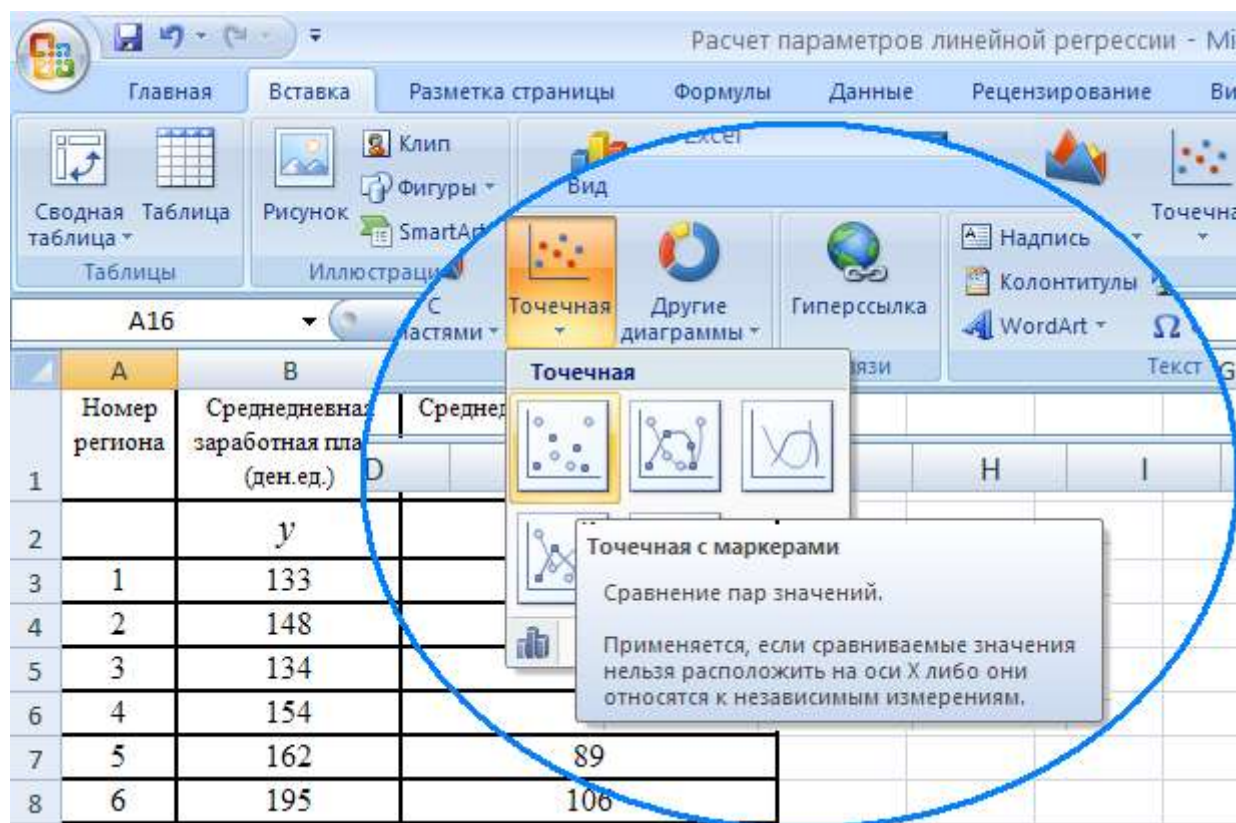


Расчет параметров линейной регрессии						
Главная Вставка Разметка страницы Формулы Данные Рецензирование						
B2 fx y						
	A	B	C	D	E	F
	Номер региона	Среднедневная заработная плата (ден.ед.)	Среднедушевой прожиточный минимум в день одного трудоспособного (ден.ед.)			
1						
2		y	x			
3	1	133	78			
4	2	148	82			
5	3	134	87			
6	4	154	79			
7	5	162	89			
8	6	195	106			
9	7	139	67			
10	8	158	88			
11	9	152	73			
12	10	162	87			
13	11	159	76			
14	12	173	115			

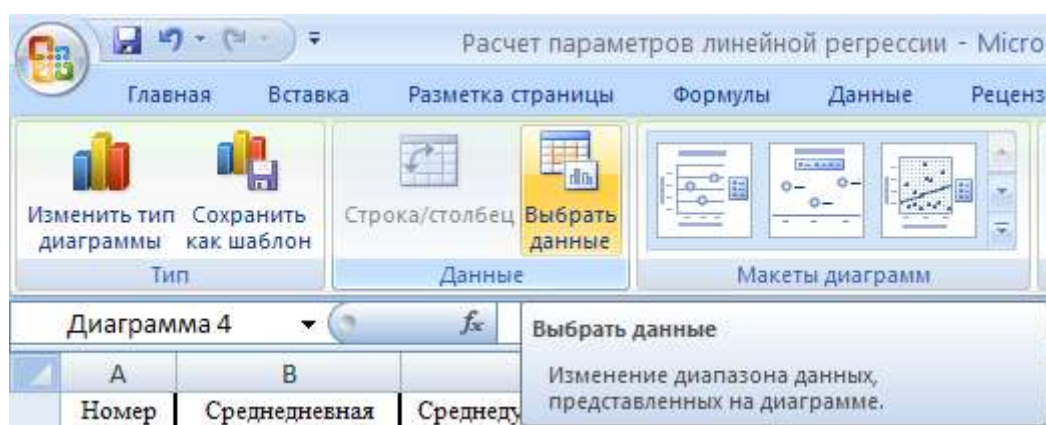
Рис.2. Исходные данные типовой задачи

Построение поля корреляции

Каждая пара наблюдений (x_i, y_i) отображается на плоскости в виде точки с соответствующими координатами. Совокупность таких точек называется полем корреляции. Чтобы построить в Excel такой график, по виду которого можно выдвинуть гипотезу о форме связи исследуемых величин, выберите на ленте интерфейса (вкладка Вставка, группа Диаграммы) тип диаграммы Точечная с маркерами (рис.3).



После нажатия на кнопку с выбранным типом диаграммы на ленте интерфейса откроется вкладка Работа с диаграммами. В группе Данные нажмите на кнопку Выбрать данные (рис.4).



В открывшемся окне выбора данных нажмите кнопку Добавить (рис.5).

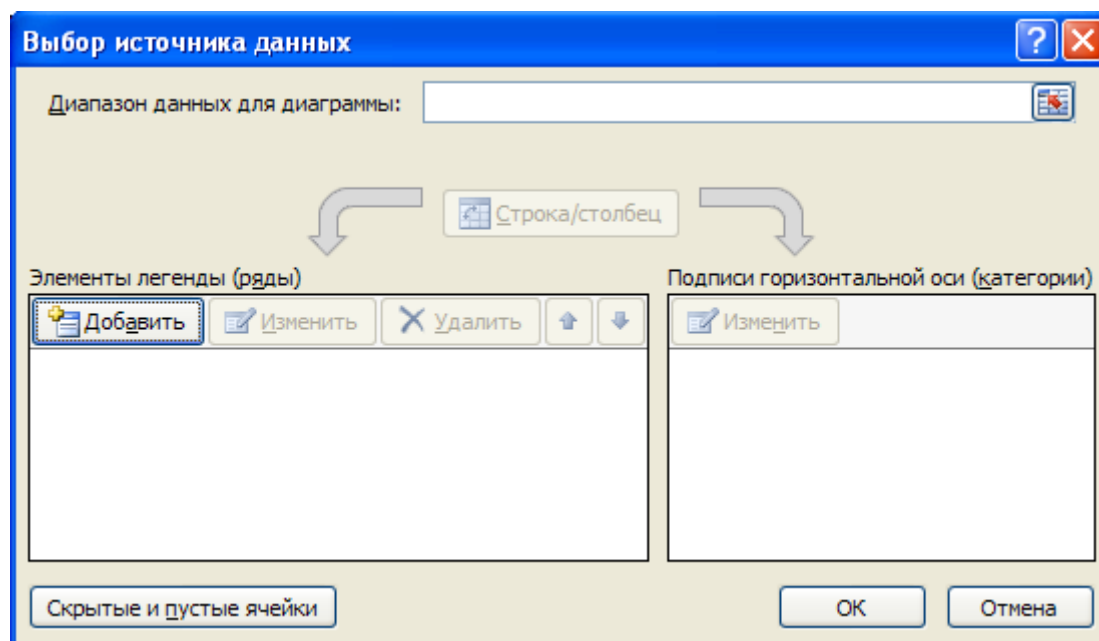


Рис.5 Диалоговое окно выбора данных

Дальше откроется окно, в которое непосредственно вводятся данные для построения диаграммы (рис.6).

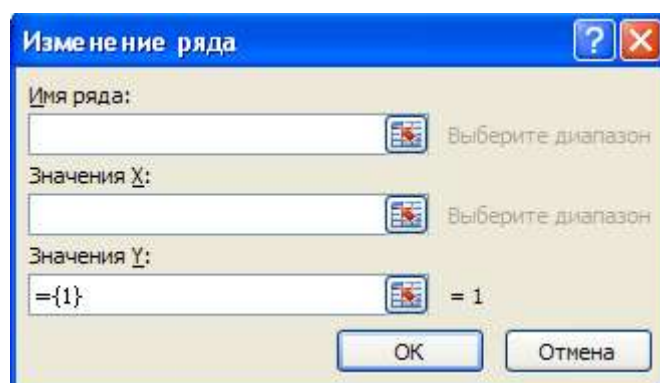


Рис.6. Окно ввода данных

В строке Значения X с помощью курсора укажите диапазон ячеек со значениями величины x (рис.7).

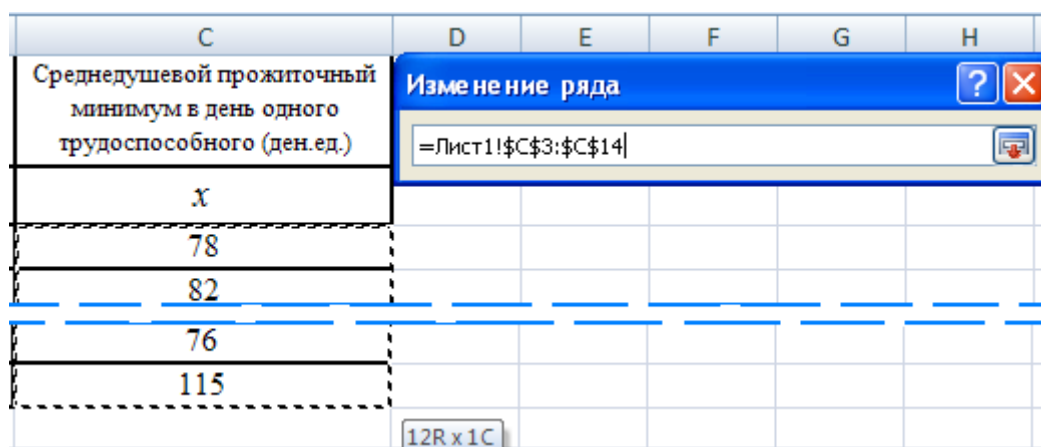


Рис.7. Ввод значений величины x

В строке Значения Y сначала удалите имеющуюся там по умолчанию запись $=\{1\}$. Затем также с помощью курсора укажите диапазон ячеек со значениями величины y. Окно для ввода данных будет выглядеть как на рис.8.

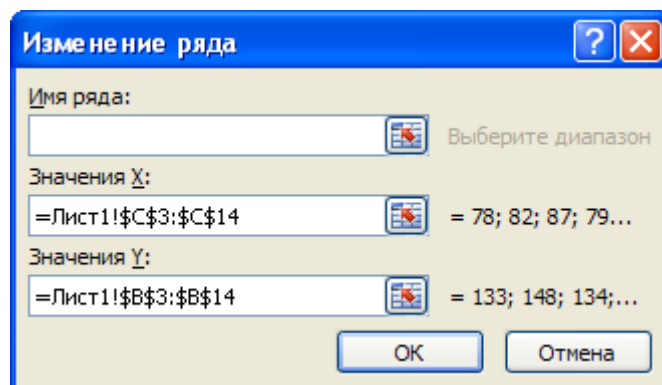


Рис.8. Окно с указанными диапазонами ячеек

После ввода данных нажмите кнопку ОК, и в появившемся окне выбора данных также нажмите кнопку ОК. Область построения диаграммы автоматически примет вид как на рис. 9.

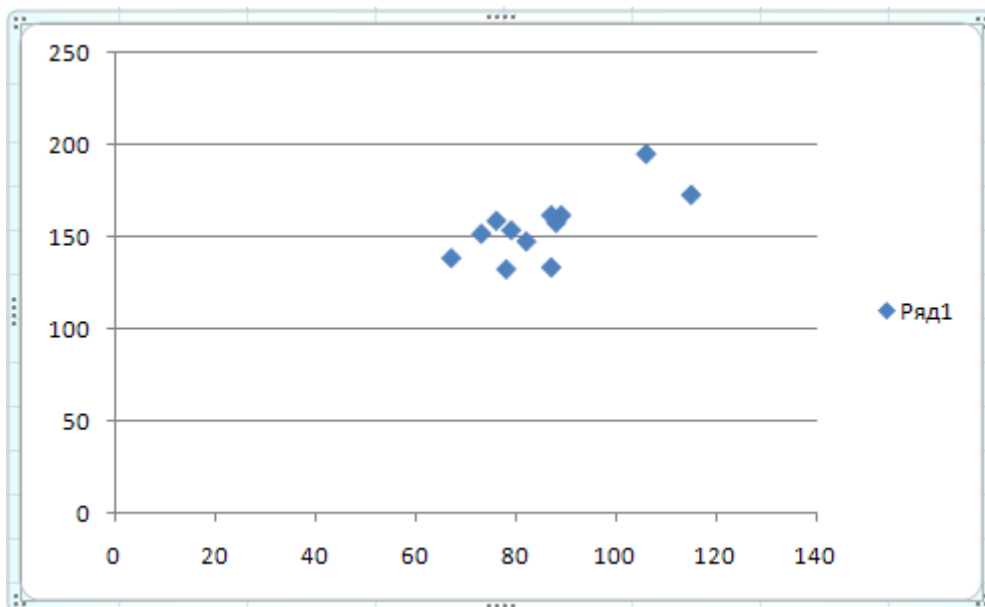


Рис.9. Область диаграммы с точками поля корреляции

На автоматически построенном поле корреляции точки расположены слишком близко друг к другу. Поэтому невозможно сделать какое-либо предположение о форме связи исследуемых величин. В таком виде диаграммой пользоваться нельзя. Чтобы лучше отобразить информацию, необходимо изменить формат осей координат. Для этого щёлкните правой кнопкой мыши сначала, например, по любому числу на оси ординат. Откроется окно меню, в котором нужно нажать кнопку **Формат оси** (рис.10). Потом откроется окно для изменения параметров формата выбранной оси координат (рис.11). Чтобы изменить численное значение начала и окончания вертикальной оси для параметров минимальное значение и максимальное значение, выберите вариант **фиксированное**, а затем введите в поле минимальное значение число 130, а в поле максимальное значение число 200 (рис. 12).



Рис. 10. Меню выбора изменения формата оси

Рис.11. Окно изменения параметров формата оси

Рис.12. Изменение параметров оси ординат

Для изменения параметров оси абсцисс нужно выполнить аналогичные действия, указав в поле минимальное значение число 60, а в поле максимальное значение число 120. В результате диаграмма преобразуется к виду как на рис. 13, и расположение точек будет более ясным.

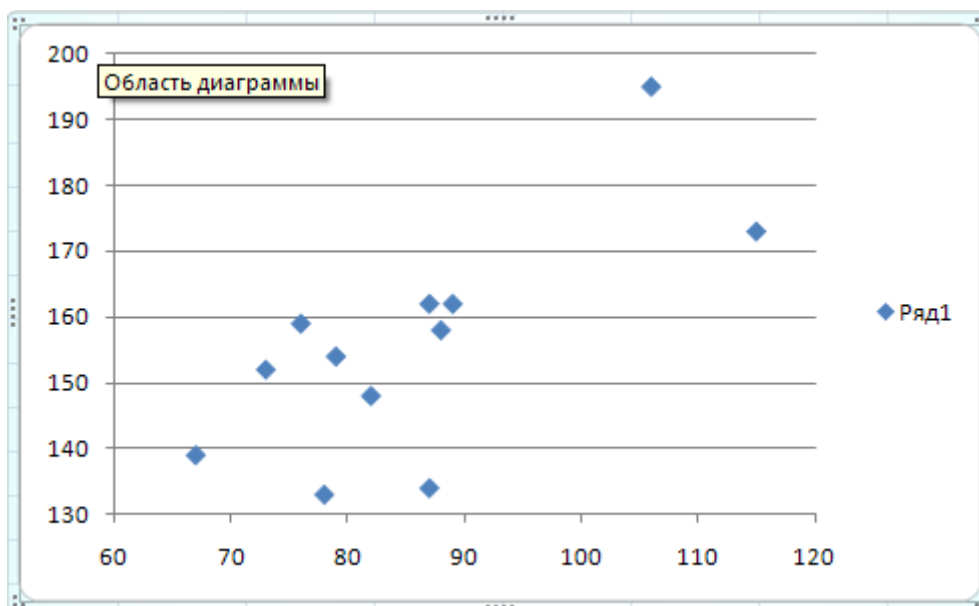


Рис.13. Поле корреляции

Построенная диаграмма в данном конкретном случае, к сожалению, не позволяет сделать однозначное предположение о форме связи исследуемых величин. Поэтому необходимо провести полное исследование с целью определения наилучшей регрессионной модели.

Расчет параметров парной линейной регрессии с помощью формул

Для расчета параметров уравнения парной линейной регрессии построим расчетную таблицу (табл.5), в которой будут отражены все промежуточные вычисления.

Таблица 5

п	x	y	xy	x^2	y^2	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$\hat{y}_{x_i}$	$(y_i - \hat{y}_{x_i})^2$	A_i
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										
11										
12										
Итого										
Среднее										
σ^2										
σ										

Поскольку вычисления удобнее проводить в Excel, то дополним таблицу исходных данных (рис. 2) необходимыми столбцами. Для правильной записи символов в Excel на ленте интерфейса выбираем вкладку Вставка, и в группе

Текст нажимаем кнопку Символ. Откроется окно, в котором можно выбрать любой нужный символ (рис.14).

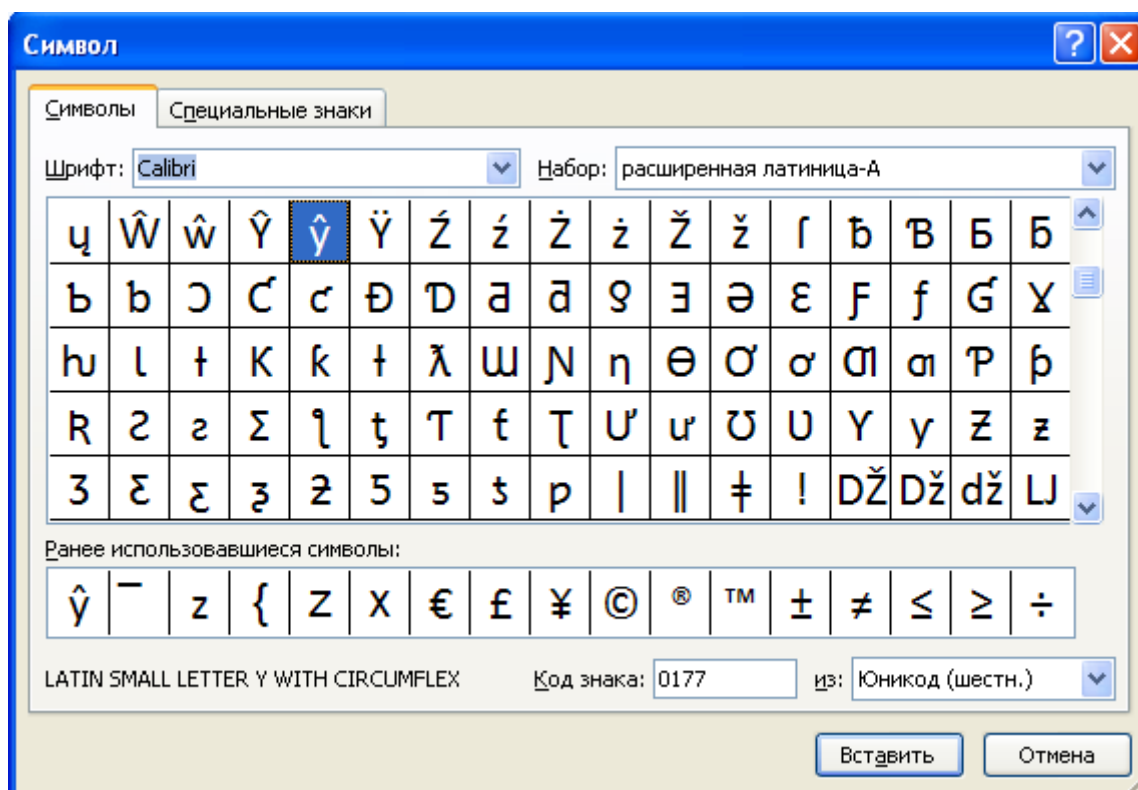


Рис. 14. Окно вставки символов

Чтобы ускорить процесс вставки нужных символов, можно в поле Код знака ввести соответствующий код. Коды для наиболее часто используемых символов приведены в табл.6.

Таблица 6

Символ	Шрифт	Набор	Код знака
$\bar{y}$	Bookshelf Symbol 7	-	66
$\hat{y}$	Arial, Calibri	расширенная латиница -А	0177
σ	Arial	греческие и коптские символы	03C3
$*^2$	Arial	дополнительная латиница - 1	00B2
$\hat{*}$	Calibri	объединенные диакр. знаки	0302
$\bar{*}$	Calibri	объединенные диакр. знаки	0304
$*_i$	Calibri	фонетические знаки	1D62

Индексы к символам можно также добавить с помощью диалогового окна Формат ячеек, которое вызывается нажатием на правую кнопку мыши (рис. 15). Перед этим убедитесь, что курсор расположен после нужного символа. В открывшемся окне выбираем параметр Видоизменение и устанавливаем флажок надстрочный (рис. 16). После нажатия кнопки ОК все вводимые символы будут отображаться в виде верхнего индекса (рис. 17). Аналогично, установив флажок подстрочный, можно вводить символы нижнего индекса. Таким способом в нижнем или в верхнем индексе можно записывать любой текст.

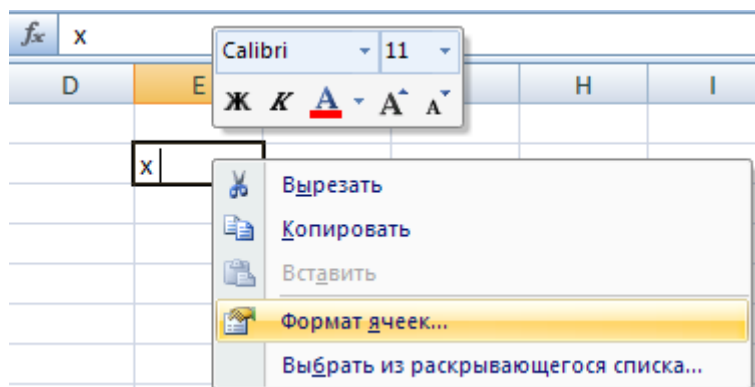


Рис. 15. Открытие окна форматирования ячеек

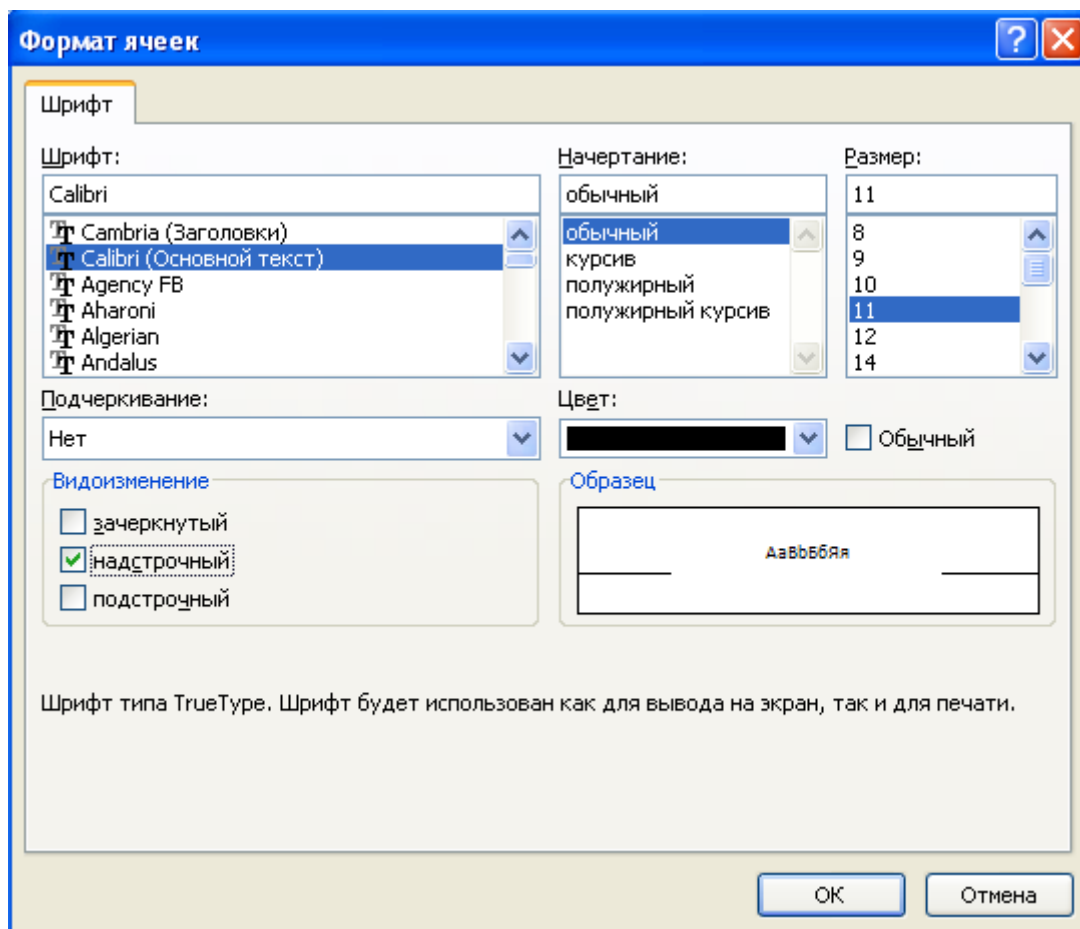


Рис. 16. Изменение параметра Видоизменение

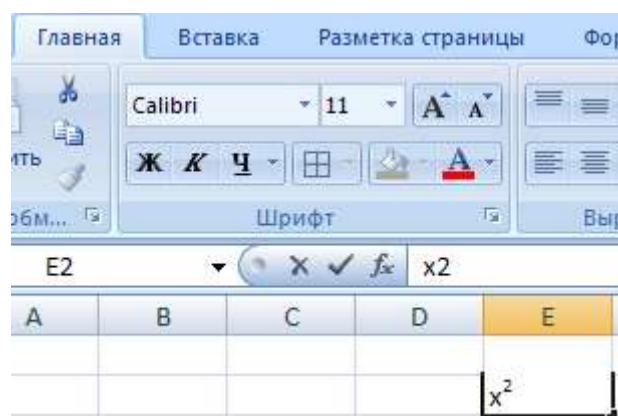


Рис. 17. Ввод верхнего индекса

Таблица для расчета параметров уравнения парной линейной регрессии в Excel с исходными данными задачи будет иметь вид как на рис.18.

	A	B	C	D	E	F	G	H	I	J	K
2		y	x	xy	x ²	y ²	(x _i - $\bar{x}$) ²	(y _i - $\bar{y}$) ²	$\hat{y}_{xi}$	(y _i - $\hat{y}_{xi}$) ²	A _i
3	1	133	78								
4	2	148	82								
5	3	134	87								
6	4	154	79								
7	5	162	89								
8	6	195	106								
9	7	139	67								
10	8	158	88								
11	9	152	73								
12	10	162	87								
13	11	159	76								
14	12	173	115								
15	Итого										
16	Среднее										
17	σ										
18	σ ²										

Рис. 18. Расчетная таблица в Excel

Далее нужно выполнить последовательно ряд действий:

1. В ячейку D3 ввести формулу =C3*B3;
2. В ячейку E3 ввести формулу =C3^2;
3. В ячейку F3 ввести формулу =B3^2;
4. Выделить диапазон ячеек D3:F3 и перетащить маркер заполнения  по ячейкам диапазона D4:F14;
5. В ячейку B15 ввести формулу =СУММ(B3:B14);
6. В ячейку B16 ввести формулу =СРЗНАЧ(B3:B14);
7. Выделить диапазон ячеек B15:B16 и перетащить маркер заполнения  по ячейкам диапазона C15:F16;
8. В ячейку G3 ввести формулу =(C3-\$C\$16)^2;
9. В ячейку H3 ввести формулу =(B3-\$B\$16)^2;
10. Выделить диапазон ячеек G3:H3 и перетащить маркер заполнения  по ячейкам диапазона G4:H14;
11. В ячейку B18 ввести формулу =F16-B16^2;
12. В ячейку B17 ввести формулу =КОРЕНЬ(B18);
13. В ячейку C18 ввести формулу =E16-C16^2;
14. В ячейку C17 ввести формулу =КОРЕНЬ(C18);

Параметры уравнения регрессии рассчитаем по формулам

$$b = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2}, a = \bar{y} - b\bar{x}.$$

Для этого выполним следующие действия:

1. В ячейке A20 запишем $b=$;
2. В ячейке A21 запишем $a=$;
3. В ячейку B20 введем формулу $=(D16-C16*B16)/(E16-C16^2)$;
4. В ячейку B21 введем формулу $=B16-B20*C16$.

Полученные таким образом значения параметров уравнения регрессии a и b используем для вычисления теоретических значений результативной переменной $\hat{y}_{x_i}$ и средней ошибки аппроксимации $\bar{A}$. Заполним нашу таблицу до конца:

1. В ячейку I3 введем формулу $=\$B\$21+\$B\$20*C3$. На ячейки B20 и B21 необходимо дать абсолютные ссылки, чтобы при автоматическом заполнении значения $\hat{y}_{x_i}$ подсчитывались верно. Для этого после указания курсором на соответствующую ячейку нужно нажать на клавиатуре кнопку F4.
2. В ячейку J3 введем формулу $=(B3-I3)^2$;
3. В ячейку K3 введем формулу $=ABS((B3-I3)/B3)*100$;
4. Выделим диапазон ячеек I3:K3 и перетащим маркер заполнения  по ячейкам диапазона I4:K14;
5. Выделим ячейку F15 и перетащим маркер заполнения  по ячейкам диапазона G15:K15;
6. В ячейку K16 введем формулу $=CPЗНАЧ(K3:K14)$.

Заполненная таблица будет иметь следующий вид (рис.19).

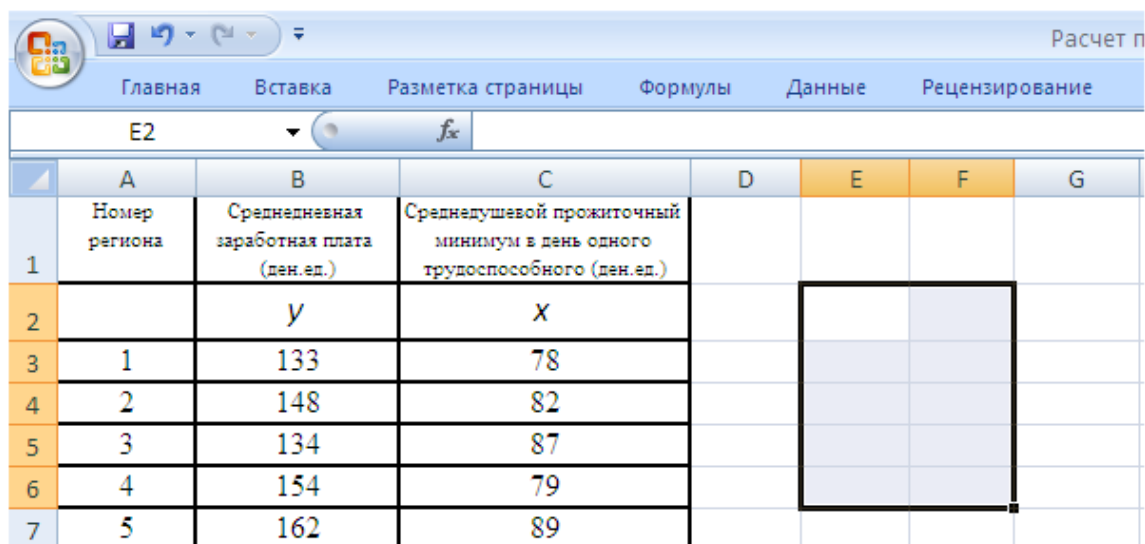
	A	B	C	D	E	F	G	H	I	J	K
2		y	x	xy	x^2	y^2	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$\hat{y}_{x_i}$	$(y_i - \hat{y}_{x_i})^2$	A_i
3	1	133	78	10374	6084	17689	57,51	517,56	148,77	248,70	11,86
4	2	148	82	12136	6724	21904	12,84	60,06	152,45	19,82	3,01
5	3	134	87	11658	7569	17956	2,01	473,06	157,05	531,48	17,20
6	4	154	79	12166	6241	23716	43,34	3,06	149,69	18,57	2,80
7	5	162	89	14418	7921	26244	11,67	39,06	158,89	9,64	1,92
8	6	195	106	20670	11236	38025	416,84	1540,56	174,54	418,52	10,49
9	7	139	67	9313	4489	19321	345,34	280,56	138,65	0,13	0,26
10	8	158	88	13904	7744	24964	5,84	5,06	157,97	0,00	0,02
11	9	152	73	11096	5329	23104	158,34	14,06	144,17	61,34	5,15
12	10	162	87	14094	7569	26244	2,01	39,06	157,05	24,46	3,05
13	11	159	76	12084	5776	25281	91,84	10,56	146,93	145,70	7,59
14	12	173	115	19895	13225	29929	865,34	297,56	182,83	96,55	5,68
15	Итого	1869	1027	161808	89907	294377	2012,917	3280,25	1869,00	1574,92	69,02
16	Среднее	155,75	85,58	13484	7492,25	24531,4					5,75
17	σ	16,53	12,95								
18	σ^2	273,35	167,74								
19											
20	b=	0,92									
21	a=	76,98									

Рис.19. Расчетная таблица с заполненными ячейками

После этого можно переходить к этапу оценки качества модели. Но сначала рассмотрим еще несколько способов расчета параметров парной линейной регрессии, которые можно реализовать с помощью средств табличного процессора Excel.

Расчет параметров парной линейной регрессии с помощью встроенных функций

Для вывода результатов регрессионной статистики с помощью встроенных функций следует выделить прямоугольную область пустых ячеек 5×2 (5 строк и 2 столбца) начиная, например, с ячейки E2 (рис. 20), или область 1×2 (одна строка, 2 столбца) для получения только оценок коэффициентов линейной регрессии.



The screenshot shows the Excel interface with the 'Формулы' (Formulas) tab active. The worksheet contains data in columns A, B, and C. A 5x2 area in columns E and F, starting from row 2, is highlighted with a black border, indicating the selected output range for the regression analysis.

	A	B	C	D	E	F	G
1	Номер региона	Среднедневная заработная плата (ден.ед.)	Среднедушевой прожиточный минимум в день одного трудоспособного (ден.ед.)				
2		Y	X				
3	1	133	78				
4	2	148	82				
5	3	134	87				
6	4	154	79				
7	5	162	89				

Рис. 20. Исходные данные и область для вывода результатов

Далее активизируйте «Мастер функций» одним из следующих способов:

- нажмите сочетание клавиш Shift+F3;
- нажмите на кнопку f_x рядом со строкой ввода, как показано на рис. 21.

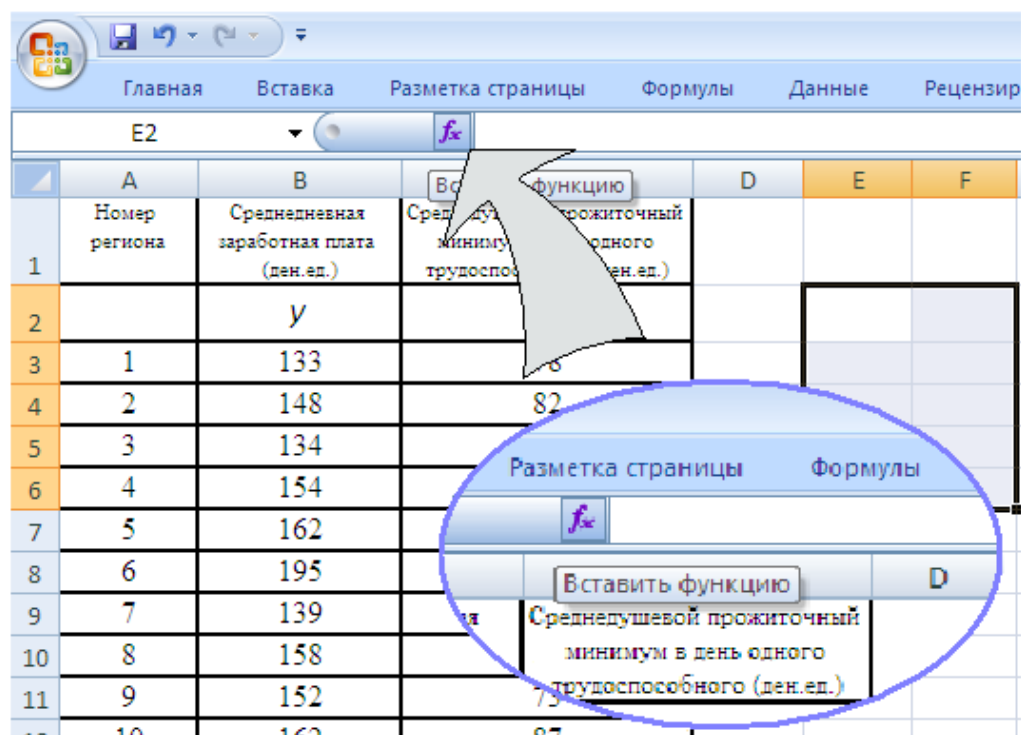


Рис. 21. Активизация инструмента «Мастер функций»

После вызова «Мастера функций» в диалоговом окне выбора функции в поле «Категория» выберите «Статистические» (рис. 22).

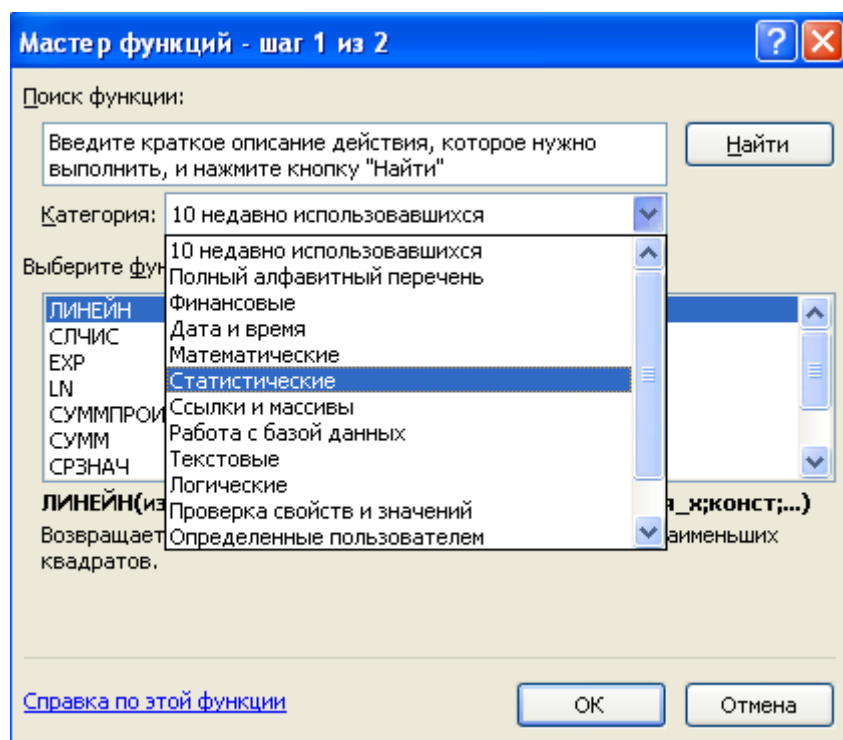


Рис. 22. Выбор категории функций

После этого среди статистических функций выберите функцию «ЛИНЕЙН».

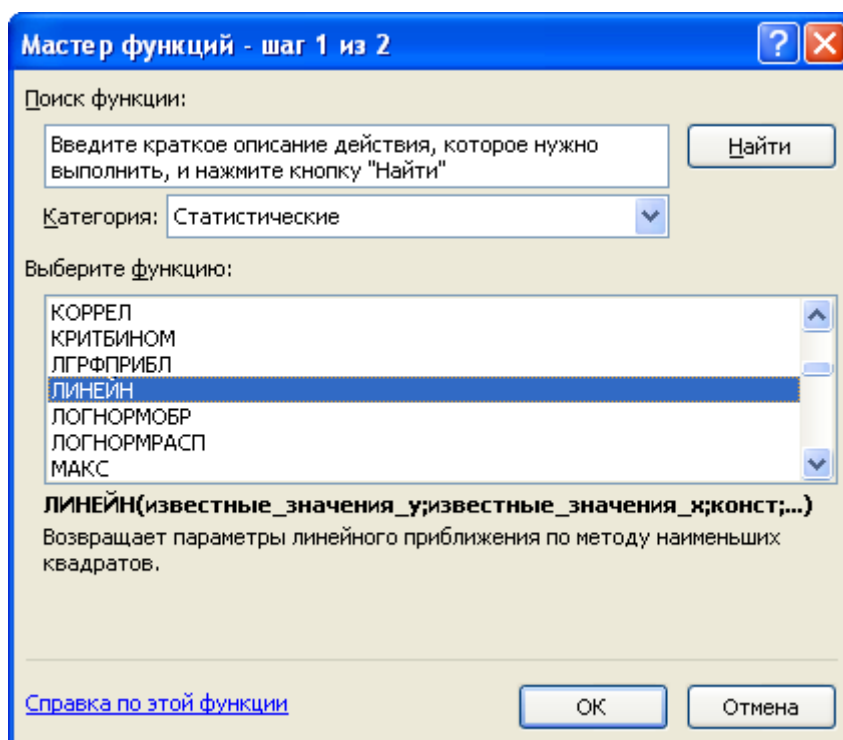


Рис.23. Выбор функции ЛИНЕЙН

При необходимости дополнительной информации о выбираемой функции можно нажать ссылку «Справка по этой функции». После нажатия на кнопку «ОК» появится диалоговое окно ввода аргументов функции (рис. 25).

в) тот же результат можно получить, если на ленте инструментов «Формулы» нажать кнопку «Другие функции», в открывшемся окне выбрать «Статистические», после чего с помощью колеса прокрутки выбрать функцию «ЛИНЕЙН» (рис.24). При нажатии на неё появляется диалоговое окно ввода аргументов функции.

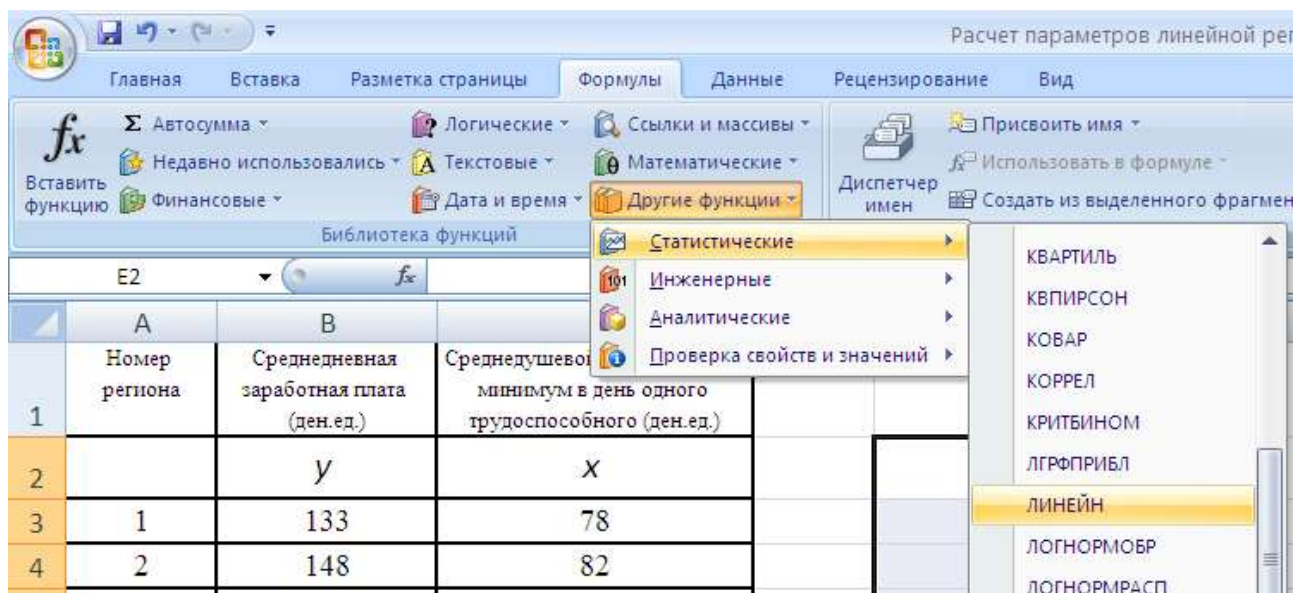


Рис. 24. Выбор функции «ЛИНЕЙН» с помощью ленты

Для заполнения аргументов функции (рис. 25) можно указать диапазон ячеек разными способами, либо набрав их адреса в соответствующих строках, либо указать мышкой прямоугольную область ячеек, отодвинув предварительно диалоговое окно в сторону.

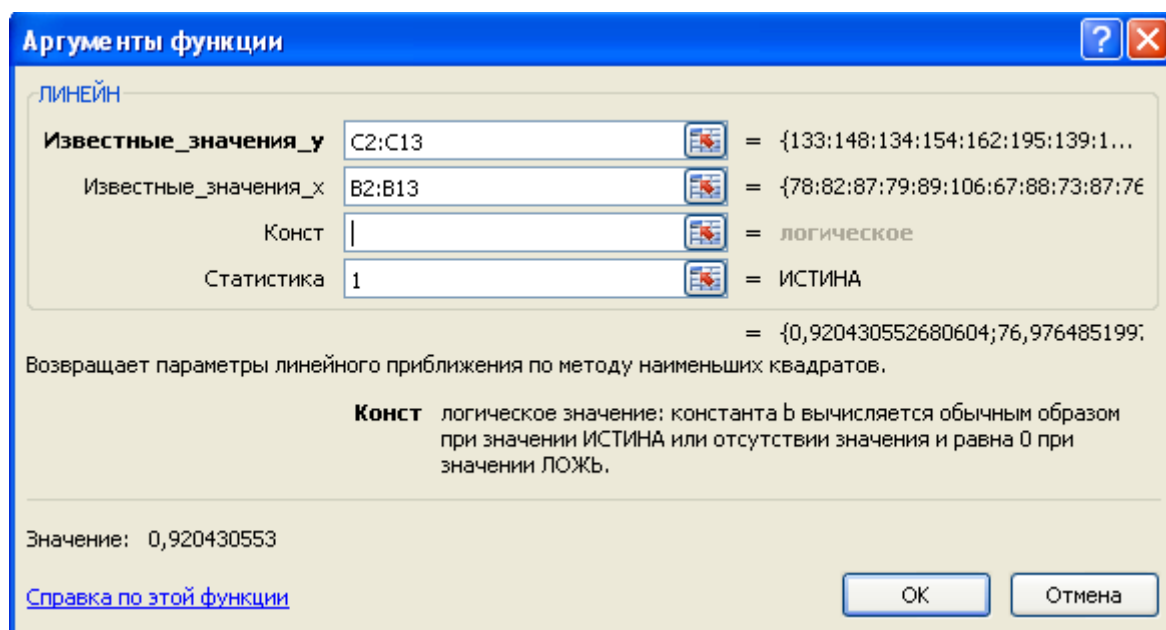


Рис.25. Диалоговое окно ввода аргументов функции «ЛИНЕЙН».

При заполнении аргументов функции в диалоговом окне появляются пояснения аргументов и констант, облегчающие их выбор.

Аргумент **Конст** необязательный. Если он имеет значение **ИСТИНА** или опущен, то коэффициент регрессии вычисляется обычным образом. Если он имеет значение **ЛОЖЬ**, то значение a полагается равным 0 и значения b подбираются таким образом, чтобы выполнялось соотношение $\hat{y} = bx$.

Если аргумент **Статистика** равен нулю, то выводятся только оценки параметров регрессии. Если он равен 1, то выводится дополнительная информация о регрессии.

При нажатии на кнопку «ОК» будет выведен только параметр наклона регрессионного уравнения. Для вывода всех параметров следует либо вместо кнопки «ОК» нажать комбинацию Ctrl+Shift+Enter (вывод массива значений), либо после нажатия кнопки «ОК» нажать F2 (редактирование формулы) и после этого нажать указанную комбинацию Ctrl+Shift+Enter.

Напомним, что для вывода массива значений необходимо до ввода формулы выделить область ячеек, соответствующую количеству выводимых значений, в нашем случае это массив ячеек 5×2.

Параметры регрессионной статистики для модели $\hat{y} = a + bx$ с помощью встроенной функции «ЛИНЕЙН» выводятся в следующем порядке (табл. 7):

Таблица 7

Значение коэффициента b	Значение коэффициента a
Среднеквадратическое отклонение σ_b	Среднеквадратическое отклонение σ_a
Коэффициент детерминации R^2	Среднеквадратическое отклонение σ_y
F – статистика Фишера	Число степеней свободы
Регрессионная сумма квадратов	Остаточная сумма квадратов

Для данных нашей задачи результат вычисления параметров регрессии представлен на рис. 26.

	A	B	C	D	E	F
	Номер региона	Среднедневная заработная плата (ден.ед.)	Среднедушевой прожиточный минимум в день одного трудоспособного (ден.ед.)			
1						
2		y	x		0,92043	76,9765
3	1	133	78		0,27972	24,2116
4	2	148	82		0,51988	12,5496
5	3	134	87		10,828	10
6	4	154	79		1705,33	1574,92

Рис.26. Результат вычисления параметров регрессии

Расчет параметров парной линейной регрессии с помощью пакета «Анализ данных»

Решение задач линейной регрессии с помощью пакета «Анализ данных» осуществляется через ленту «Данные», группа «Анализ», пункт «Анализ данных» (рис. 27).

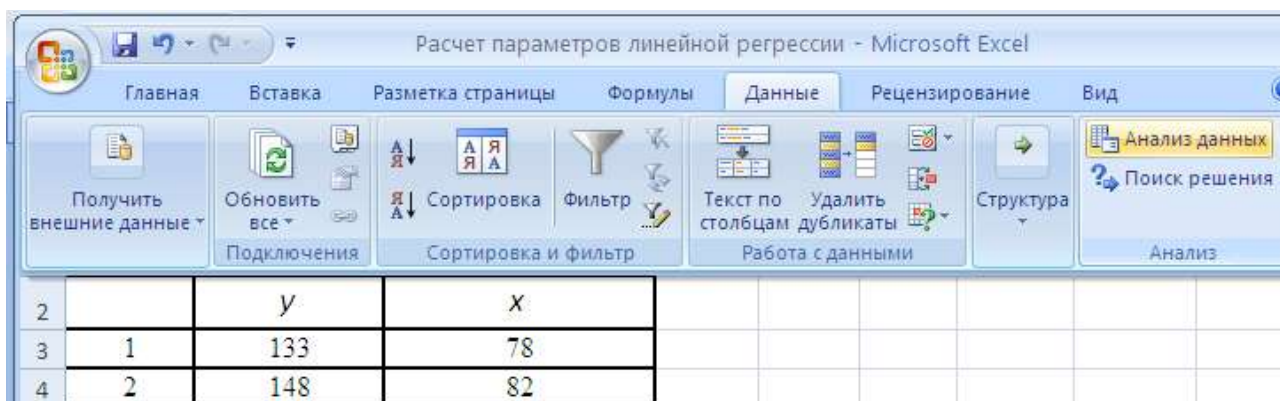


Рис.27. Выбор инструмента «Анализ данных»

Если этот пункт отсутствует на ленте «Данные», то необходимо активировать эту надстройку, нажав последовательно кнопки «Office», «Параметры Excel», «Надстройки», «Перейти», поставить галочку напротив пункта Пакет анализа и нажать кнопку ОК (рис.28).

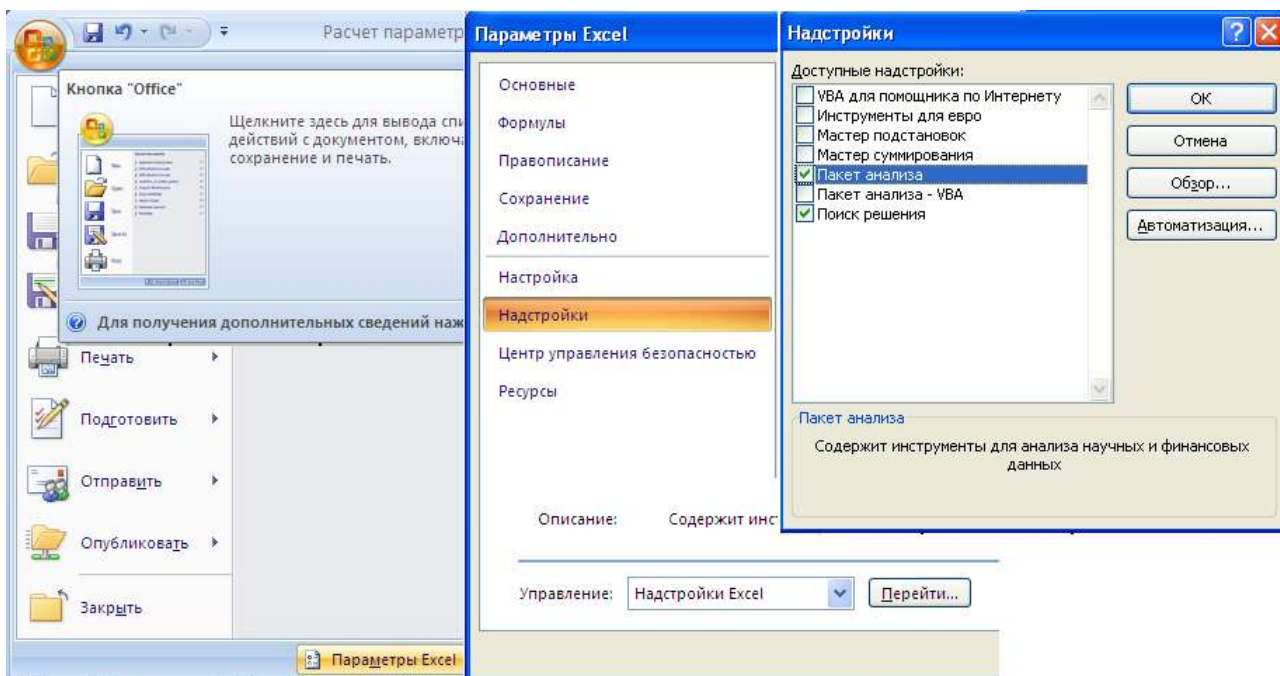


Рис. 28. Активизация инструмента «Пакет анализа»

Исходные данные размещены на первом листе в столбцах А, В, С (рис.2). Поместите курсор в свободную ячейку и вызовите пункт «Анализ данных»

группы «Анализ» ленты «Данные». В открывшемся диалогом окне (рис.29) выберите пункт Регрессия.

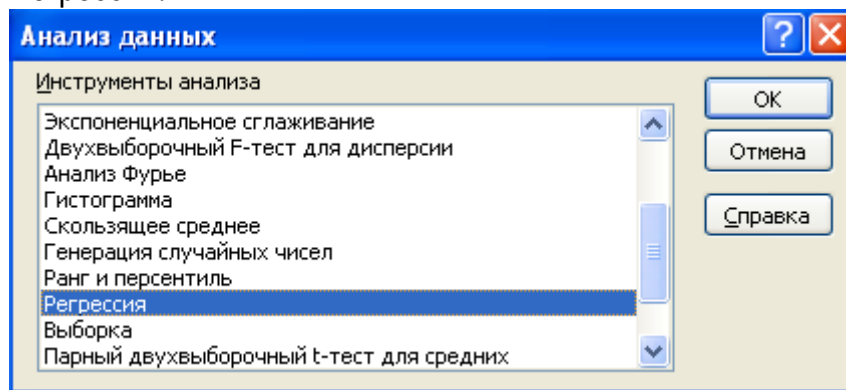


Рис.29. Выбор вида анализа данных

Диалоговое окно «Регрессия» содержит поля для ввода входных данных и основных параметров вывода (рис.30).

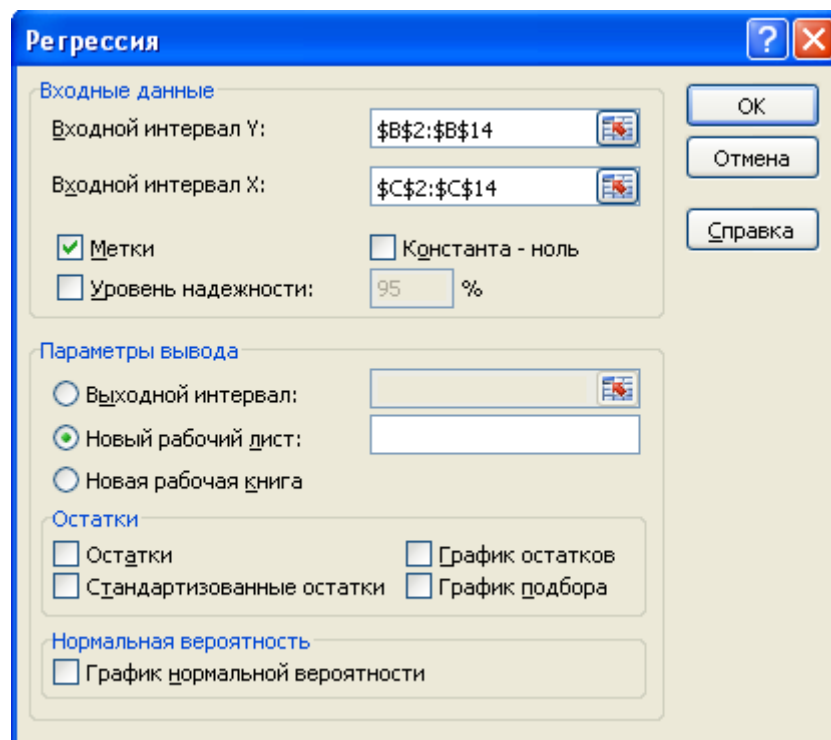


Рис.30. Диалоговое окно ввода параметров

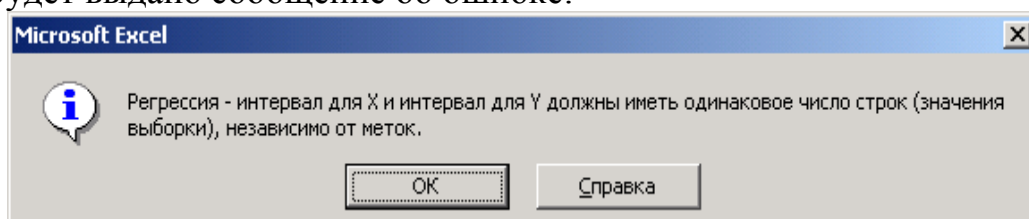
Если поле **Метки** включено, в качестве входного интервала Y следует указать не только адреса ячеек, содержащих значения анализируемой функции, но и адрес клетки B2, содержащей подпись к столбцу значений функции. То же относится и к полю ввода **Входной интервал X**.

Если включен флажок в поле **Уровень надежности**, можно определить значение доверительной вероятности, для которой будут определены Верхняя и Нижняя границы доверительного интервала дополнительно к уровню 95%, применяемому по умолчанию.

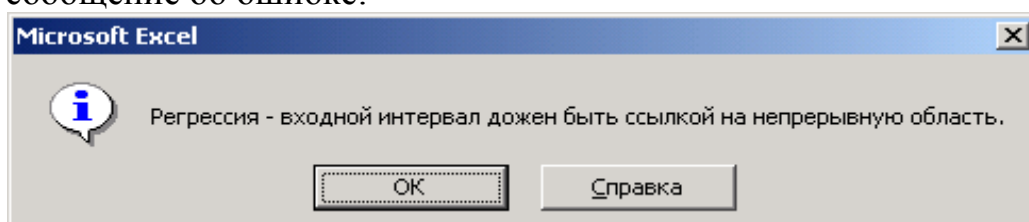
Параметры вывода диалогового окна определяют область вывода: «отдельная книга», «новый рабочий лист» либо размещение результата в заданной области рабочего листа (указывается левая верхняя ячейка).

С помощью инструмента анализ данных «Регрессия» помимо результатов регрессионной статистики, дисперсионного анализа и доверительных интервалов можно получать остатки и графики подбора линий регрессии, остатков и нормальной вероятности, включив соответствующие метки в нижней части диалогового окна «Регрессии». Нажатие кнопки «ОК» завершает ввод данных и параметров вывода.

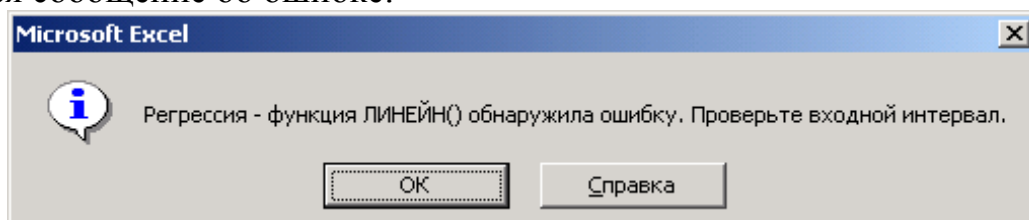
В случае, если для функции и параметров указано разное количество строк, будет выдано сообщение об ошибке:



Если во множественной регрессии столбцы факторов не смежные, будет выдано сообщение об ошибке:



Если во множественной регрессии факторы влияния линейно зависимы, появится сообщение об ошибке:



В любом случае появления сообщения об ошибке нажмите «ОК», вернитесь к диалоговому окну ввода параметров и проверьте правильность ввода данных.

В случае отсутствия ошибок результаты решения будут представлены в указанной области вывода.

По окончании расчета на рабочий лист выводится три группы результатов (рис. 31).

Первая группа Регрессионная статистика включает в свой состав:

- Множественный R - коэффициент множественной корреляции;
- R -квадрат - множественный коэффициент детерминации;
- Нормированный R -квадрат - скорректированный коэффициент детерминации, который вводится для устранения эффекта, связанного с ростом величины R -квадрат при возрастании числа факторов;
- Стандартная ошибка – стандартная ошибка регрессии;

- Наблюдения - количество наблюдений.

Вторая группа результатов Дисперсионный анализ, где использован ряд общепринятых сокращений, содержит:

- df – число степеней свободы (degree of freedom);
- SS – сумма квадратов отклонений (Sum of squares);
- MS – средний квадрат отклонения (Mean square);
- F – статистика Фишера (наблюдаемое значение критерия Фишера);
- *Значимость F* – критическое значение квантиля распределения Фишера, на котором отвергается нулевая гипотеза отсутствия влияния фактора.

Построчно в таблице выводятся показатели, характеризующие изменчивость: присущую модели и случайную.

- *Регрессия* – здесь выводятся характеристики, связанные с закономерной изменчивостью: сумма квадратов отклонений между группами SS , соответствующее число степеней свободы df , на основании количества которых определяется «Значимость F » и частное от этих величин - средний квадрат отклонений MS . В этой же строке выведен собственно результат анализа: F —отношение дисперсий для исходных данных.
- *Остаток* – в этой строке представлены показатели, характеризующие действие случайных факторов - те же самые, что и для предыдущей строки, только без окончательных результатов анализа.
- *Итого* – представлены суммы квадратов отклонений от среднего SS и количество степеней свободы по значениям регрессии и остатка. В данном случае SS - характеристика полной изменчивости.

Следующая, третья группа результатов включает в свой состав значения коэффициентов уравнения регрессии, а также статистики, на основании которых проверяется значимость влияния фактора для каждого коэффициента, включенного в модель.

- *Коэффициенты* – значения коэффициентов уравнения линейной регрессии;
- *Стандартная ошибка* – стандартная ошибка коэффициентов;
- *t-статистика* – значение статистики критерия, на основании которого определяется уровень значимости отклонения гипотезы равенства коэффициентов нулю (*P-значение*);
- *P-значение* – уровень значимости, на котором отвергается гипотеза равенства коэффициентов нулю;
- *Нижние 95%* - нижняя граница доверительного интервала, в котором находится значение коэффициентов генеральной совокупности;
- *Верхние 95%* - верхняя граница доверительного интервала, в котором находится значение коэффициентов генеральной совокупности;
- *Нижние 99%* - нижняя граница доверительного интервала, в котором находится значение коэффициентов генеральной совокупности (значение задается при определении параметров анализа);
- *Верхние 99%* - верхняя граница доверительного интервала, в котором находится значение коэффициентов генеральной совокупности (значение задается при определении параметров анализа).

Если в диалоговом окне ввода параметров установлены флажки вывода остатков и стандартизованных остатков, то они также выводятся в таблице, где для каждого значения ряда выводится предсказанное значение, с которым сопоставляется остаток, представляющий собой разность между прогнозным и реальным значением ряда.

Кроме вывода табличной информации, если установлены соответствующие флажки в окне ввода, в области вывода строятся графики остатков, подбора для каждого фактора в отдельности и график нормального распределения, построенный по размещенным в таблице вероятности перцентилям. Для рассматриваемой задачи графики приведены на рисунке 32.

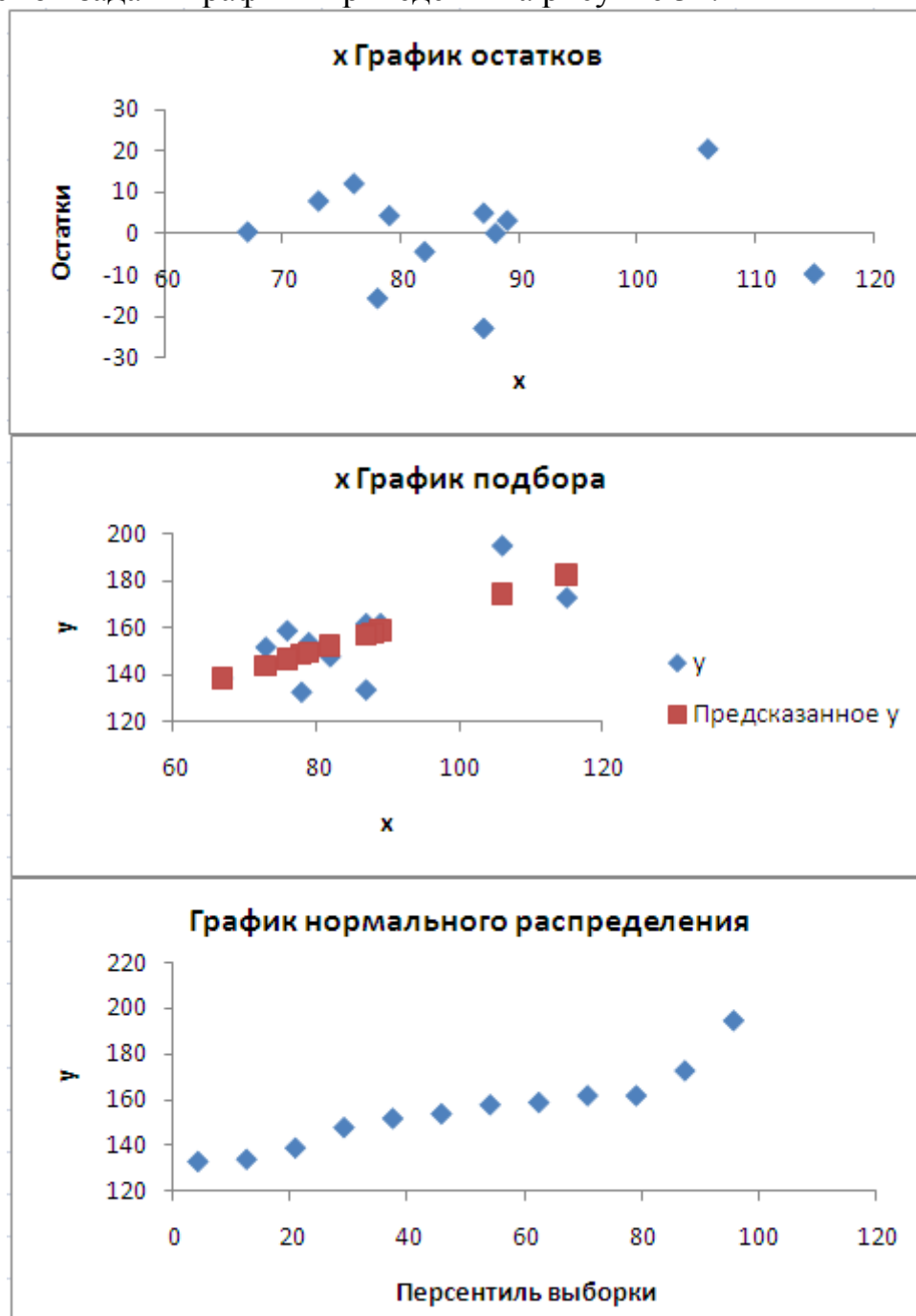


Рис.32. Графики результатов решения задачи регрессионного анализа

Расчет параметров парной линейной регрессии с помощью диаграмм

С помощью диаграммы можно быстро получить значения коэффициентов уравнения парной линейной регрессии. Диаграмма строится так же, как описано выше в разделе «Построение поля корреляции», только нужно в столбце В разместить значения переменной x , а в столбце С – значения переменной y . Другой способ построения диаграммы выглядит так. Выделите курсором значения переменных x и y , затем на вкладке Вставка в группе Диаграммы выберите тип диаграммы Точечная и затем подтип диаграммы Точечная с маркерами (рис.33).

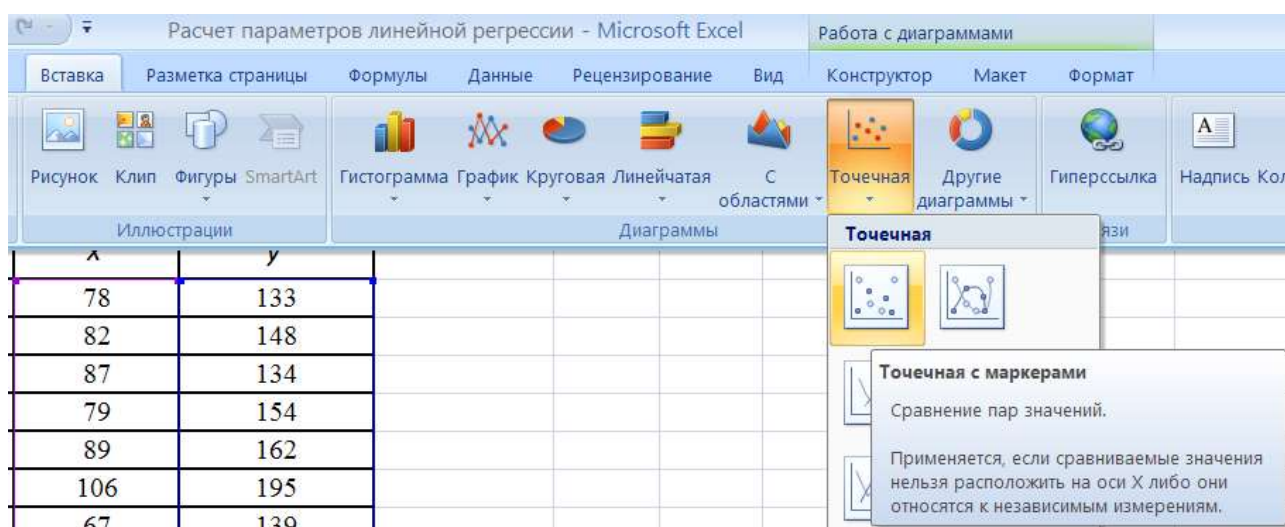


Рис.33. Выбор нужного типа диаграммы

Получится такая же диаграмма как на рис.9. Далее на вкладке Конструктор в группе Макеты диаграммы выберите макет №9 (рис.34).

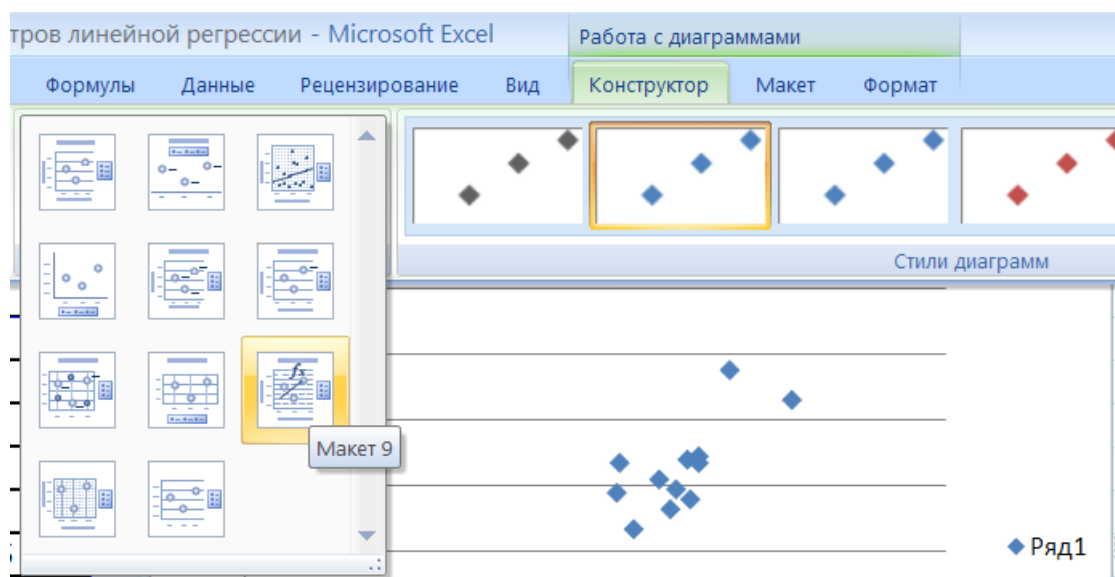


Рис.34. Выбор макета диаграммы

В результате будем иметь диаграмму, на которой уже есть готовое уравнение регрессии и коэффициент детерминации, по которому можно оценить качество модели (рис.35).



Рис.35. Диаграмма по макету №9

Расчет параметров парной нелинейной регрессии

а) *Степенная регрессия $\hat{y} = ax^b$;*

Параметры уравнения степенной регрессии можно рассчитать, предварительно линеаризовав уравнение с помощью логарифмирования, как было сказано выше. Но можно получить тот же результат быстрее, используя диаграмму.

Диаграмма с линейной регрессией по макету №9 после изменения параметров осей координат и ввода названий приобретает следующий вид (рис.36).

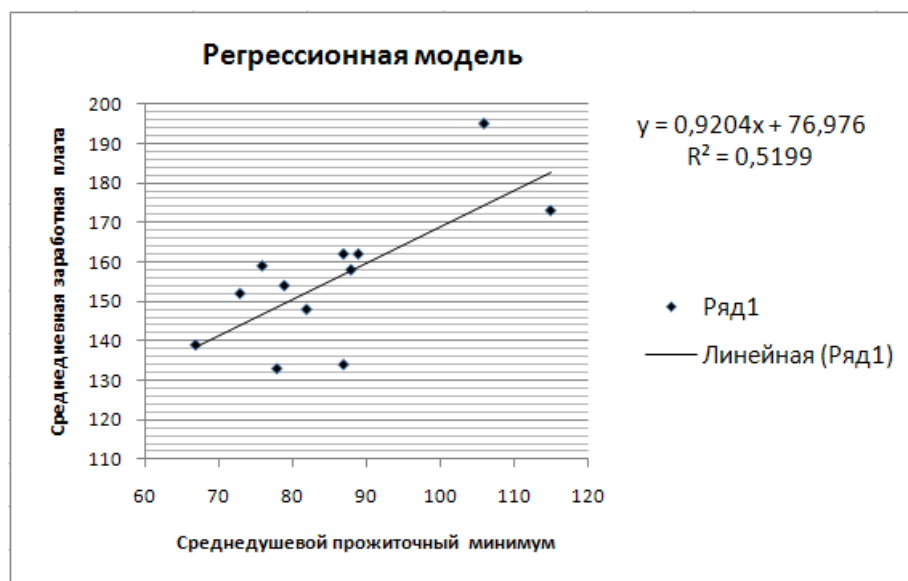


Рис.36. Линейная регрессионная модель

Теперь, чтобы получить степенную регрессионную модель, нужно на диаграмме щёлкнуть правой кнопкой мыши по линии и в открывшемся окне выбрать Формат линии тренда (рис.37).

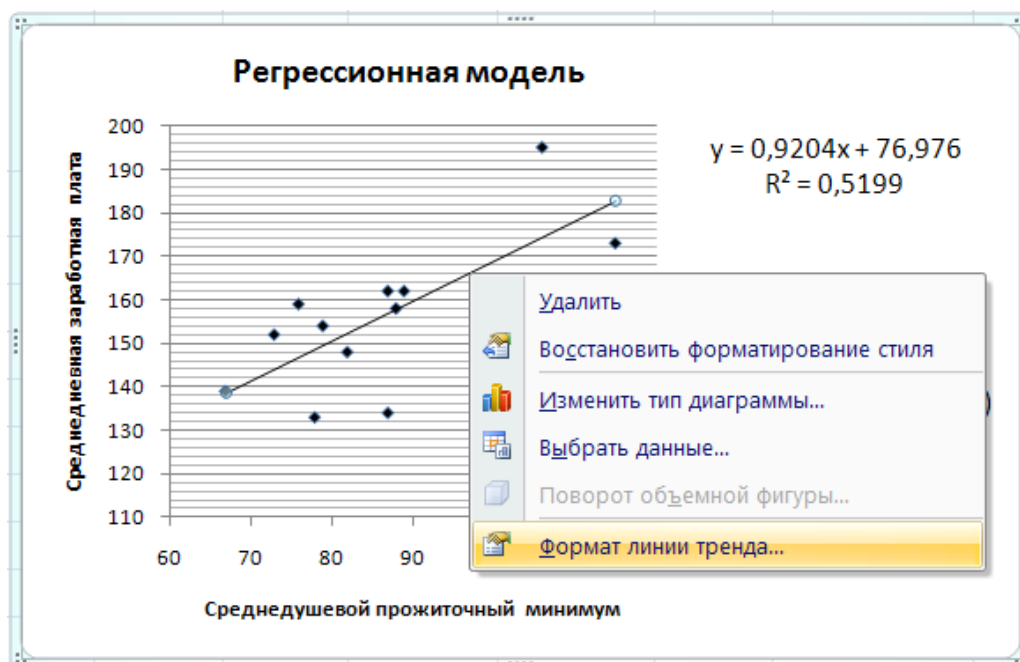


Рис.37. Выбор формата линии тренда

В открывшемся диалоговом окне (рис.38) в разделе Параметры линии тренда нужно выбрать степенную и нажать кнопку Заккрыть.

Формат линии тренда

Параметры линии тренда

Цвет линии

Тип линии

Тень

Построение линии тренда (аппроксимация и сглаживание)

- ☐ Экспоненциальная
- ☐ Линейная
- ☐ Логарифмическая
- ☐ Полиномиальная Степень: 2
- ☒ Степенная
- ☐ Линейная фильтрация Точка: 2

Название аппроксимирующей (сглаженной) кривой

- ☒ автоматическое: Степенная (Ряд1)
- ☐ другое:

Прогноз

вперед на: 0,0 периодов

назад на: 0,0 периодов

☐ пересечение кривой с осью Y в точке: 0,0

☒ показывать уравнение на диаграмме

☒ понести на диаграмму величину достоверности аппроксимации (R^2)

Заккрыть

Рис.38. Выбор вида линии тренда

Диаграмма автоматически изменится. На графике отобразится нелинейная (степенная) линия регрессии, в области диаграммы появится уравнение степенной регрессии и значение коэффициента детерминации R^2 (рис.39).

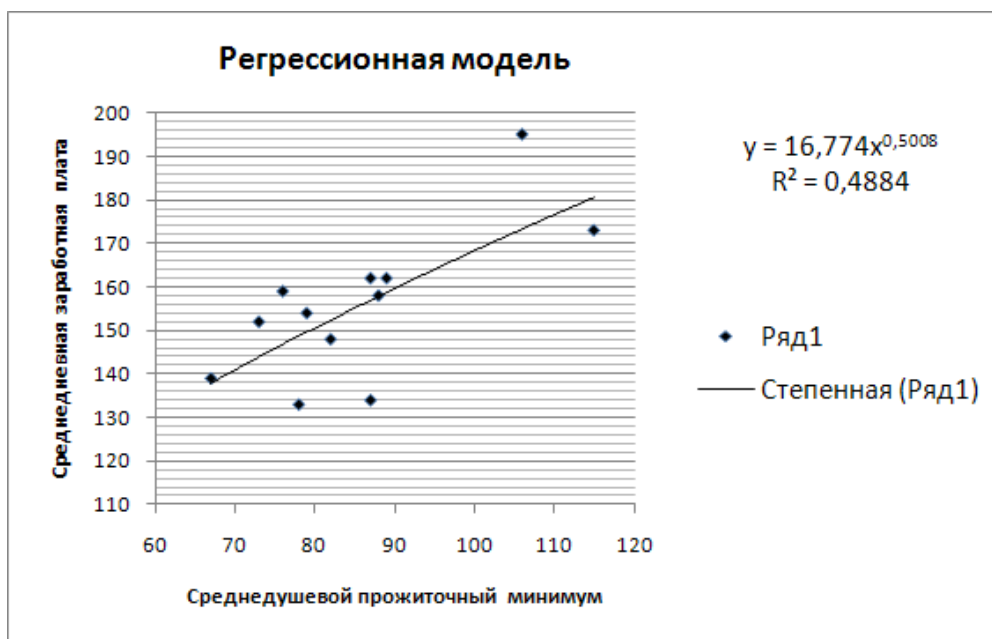


Рис. 39. Диаграмма степенной регрессии

Используя готовое уравнение, можно легко рассчитать среднюю ошибку аппроксимации $\bar{A}$ и индекс корреляции ρ_{xy} (рис. 40).

	A	B	C	D	E	F	G	H	I
	Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного (ден.ед.)	Среднедневная заработная плата (ден.ед.)						
1									
2		x	y	$\hat{y}$	$y - \hat{y}$	A_i	$y_i - \bar{y}$	$\bar{A} =$	5,78%
3	1	78	133	148,66	-15,66	11,78%	-22,75	$\rho_{xy} =$	0,717
4	2	82	148	152,43	-4,43	2,99%	-7,75		
5	3	87	134	157,02	-23,02	17,18%	-21,75		
6	4	79	154	149,61	4,39	2,85%	-1,75		
7	5	89	162	158,81	3,19	1,97%	6,25		
8	6	106	195	173,34	21,66	11,11%	39,25		
9	7	67	139	137,76	1,24	0,89%	-16,75		
10	8	88	158	157,92	0,08	0,05%	2,25		
11	9	73	152	143,81	8,19	5,39%	-3,75		
12	10	87	162	157,02	4,98	3,08%	6,25		
13	11	76	159	146,74	12,26	7,71%	3,25		
14	12	115	173	180,57	-7,57	4,37%	17,25		
15			155,75						

Рис.40. Расчёт параметров степенной регрессии

б) Показательная регрессия $\hat{y} = ab^x$;

Параметры уравнения показательной модели легко получить, приведя его к линейному виду путем логарифмирования

$$\ln y = \ln a + x \ln b,$$

$$Y = A + Bx,$$

здесь: $Y = \ln y$, $B = \ln b$, $A = \ln a$.

Далее, скопировав исходные данные на новый лист, заполняем следующую таблицу (рис. 41).

	A	B	C	D	E	F	G	H
1	Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного (ден.ед.)	Среднедневная заработная плата (ден.ед.)					
2		x	y	$\ln y$	$\hat{y}$	$y - \hat{y}$	A_i	$y_i - \bar{y}$

Рис. 41. Таблица для расчёта параметров показательной регрессии

Для этого выполняем следующие действия:

1. В ячейку D3 вводим формулу $=\text{LN}(C3)$;
2. Выделяем ячейку D3 и протягиваем маркер заполнения  по ячейкам диапазона D4:D14;
3. Строим график по макету диаграмм №9, указав в качестве значений y значения $\ln y$ из диапазона D3:D14 (рис.42).

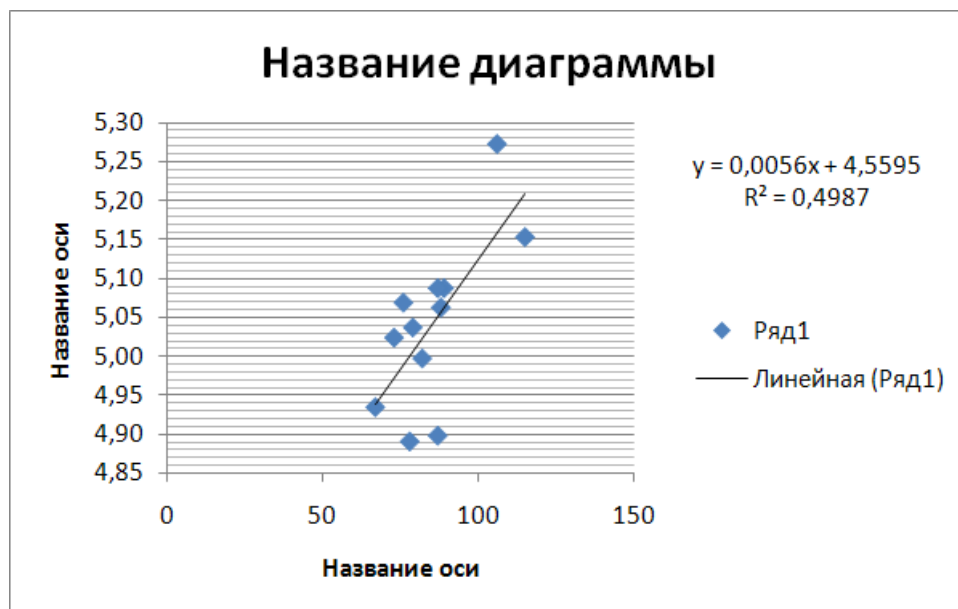


Рис. 42. График линейризованной модели $Y = A + Bx$

4. По значениям $B = \ln b$, $A = \ln a$, взятым с диаграммы (рис. 42), рассчитываем значения параметров a и b (рис. 43):

	J	K	L	M
2	lna=	4,5595	a=	95,54
3	lnb=	0,0056	b=	1,01
4	$\bar{A}$ =	5,89%		
5	ρ_{xy} =	0,72		

Рис.43. Расчёт параметров уравнения показательной регрессии

- 4.1. В ячейку M2 вводим формулу =EXP(K2);
- 4.2. В ячейку M3 вводим формулу =EXP(K3);
5. Для расчёта теоретических значений $\hat{y}$ в ячейку E3 вводим формулу =M\$2*(M\$3^B3);
6. Для расчёта отклонений теоретических значений от наблюдаемых в ячейку F3 вводим формулу =C3-E3;
7. Для расчёта относительных отклонений в ячейку G3 вводим формулу =ABS(F3/C3), формат ячейки G3 назначаем Процентный;
8. В ячейку C15 вводим формулу =СРЗНАЧ(C3:C14);
9. В ячейку H3 вводим формулу =C3-\$C\$15;
10. Выделяем ячейки E3:H3 и протягиваем маркер заполнения  по ячейкам диапазона E4:H14;
11. Для расчёта среднего относительного отклонения в ячейку K4 вводим формулу =СРЗНАЧ(G3:G14);
12. Для расчёта индекса корреляции показательной регрессии в ячейку K5 вводим формулу =КОРЕНЬ(1-СУММКВ(F3:F14)/СУММКВ(H3:H14));

в) *Полулогарифмическая регрессия* $\hat{y} = a + b \cdot \ln x$;

Параметры уравнения полулогарифмической регрессионной модели можно получить аналогично параметрам степенной регрессии.

Сначала строим график исходных данных по макету диаграмм №9, в котором меняем линию тренда на логарифмическую (рис. 44).

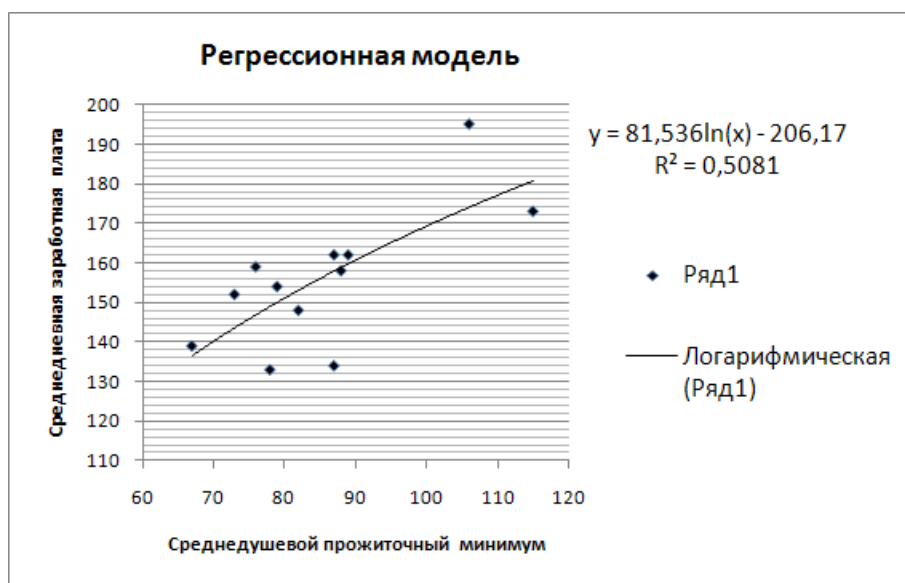


Рис. 44. Диаграмма полулогарифмической регрессии

Затем, используя готовое уравнение, рассчитываем среднюю ошибку аппроксимации $\bar{A}$ и индекс корреляции ρ_{xy} (рис. 45).

	A	B	C	D	E	F	G	H	I
1	Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного (ден.ед.)	Среднедневная заработная плата (ден.ед.)						
2		x	y	$\hat{y}$	$y-\hat{y}$	A_i	$y_i-\bar{y}$	$\bar{A}=$	5,86%
3	1	78	133	149,06	-16,06	12,07%	-22,75	$\rho_{xy}=$	0,713
4	2	82	148	153,14	-5,14	3,47%	-7,75		
5	3	87	134	157,96	-23,96	17,88%	-21,75		
6	4	79	154	150,10	3,90	2,53%	-1,75		
7	5	89	162	159,82	2,18	1,35%	6,25		
8	6	106	195	174,07	20,93	10,73%	39,25		
9	7	67	139	136,66	2,34	1,68%	-16,75		
10	8	88	158	158,89	-0,89	0,57%	2,25		
11	9	73	152	143,66	8,34	5,49%	-3,75		
12	10	87	162	157,96	4,04	2,49%	6,25		
13	11	76	159	146,94	12,06	7,58%	3,25		
14	12	115	173	180,71	-7,71	4,46%	17,25		
15			155,75						

Рис.45. Расчёт параметров уравнения полулогарифмической регрессии

г) *Гиперболическая регрессия* $\hat{y} = a + \frac{b}{x}$.

Уравнение гиперболы $\hat{y} = a + \frac{b}{x}$ линеаризуется заменой $z = \frac{1}{x}$. В результате также получается линейное уравнение, к которому применим МНК,

$$y = a + bz.$$

Расчёт параметров производим аналогично расчёту параметров для показательной регрессии. Сначала, скопировав исходные данные на новый лист, заполняем следующую таблицу (рис. 46).

	A	B	C	D	E	F	G	H
1	Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного (ден.ед.)		Среднедневная заработная плата (ден.ед.)				
2		x	1/x	y	$\hat{y}$	$y-\hat{y}$	A_i	$y_i-\bar{y}$

Рис. 46. Таблица для расчёта параметров гиперболической регрессии

Для этого выполняем следующие действия:

1. В ячейку C3 вводим формулу =1/B3;

2. Выделяем ячейку C3 и протягиваем маркер заполнения  по ячейкам диапазона C4:C14;
3. Строим график по макету диаграмм №9, указав в качестве значений X значения $1/x$ из диапазона C3:C14 (рис.47);

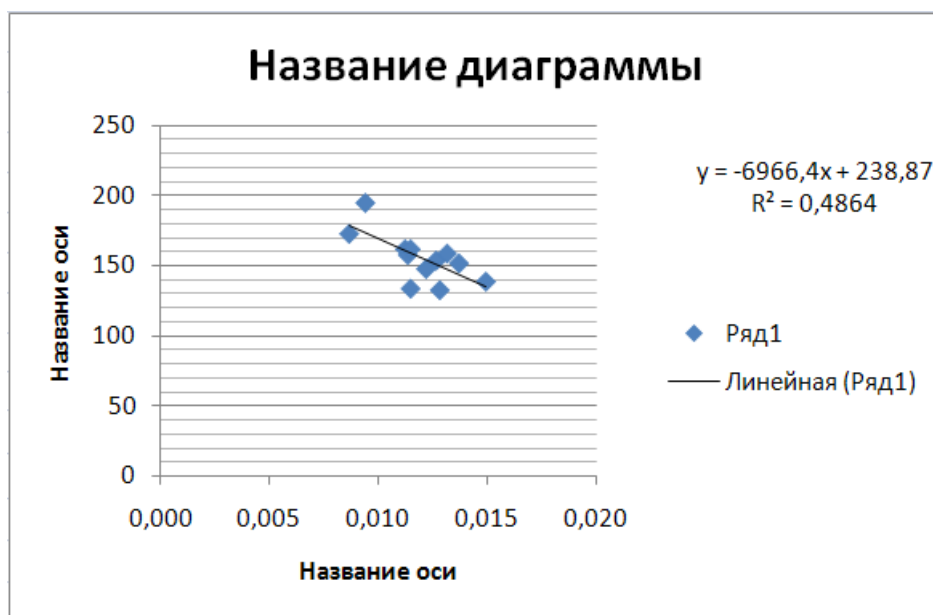


Рис. 47. График линеаризованной модели $y = a + bz$

Выполнив действия, аналогичные тем, которые выполнялись при расчёте параметров показательной модели, получим искомые значения (рис.48).

	I	J	K
2		a=	238,87
3		b=	-6966,4
4		$\bar{A}$ =	5,94%
5		ρ_{xy} =	0,697

Рис.48. Расчёт параметров уравнения гиперболической регрессии

Расчет доверительных интервалов для параметров уравнения регрессии

После расчета параметров уравнения парной линейной регрессии оценим их статистическую значимость. Сначала рассчитаем стандартные ошибки: m_b и m_a , используя встроенные функции Excel: $m_b = 0,255$, $m_a = 22,102$. Затем вычислим фактические значения t -критерия:

$$t_b = \frac{0,9204}{0,255} = 3,609, \quad t_a = \frac{76,976}{22,102} = 3,483.$$

Табличное значение t -критерия Стьюдента при уровне значимости $\alpha = 0,05$ и числе степеней свободы $df = 12 - 2 = 10$ рассчитаем с помощью

встроенной функции СТЬЮДЕНТРАСПОБР(0,05;10): $t_{\text{табл}} = 2,23$. Поскольку фактические значения t -критерия для обоих параметров больше табличного, делаем вывод о статистической значимости рассчитанных параметров.

Для расчета доверительного интервала находим предельную ошибку для каждого показателя:

$$\Delta_a = t_{\text{табл}} \cdot m_a = 49,25, \quad \Delta_b = t_{\text{табл}} \cdot m_b = 0,57.$$

Затем определяем границы доверительных интервалов как $a \pm t_{\text{табл}} \cdot m_a$, $b \pm t_{\text{табл}} \cdot m_b$. Получаем следующие значения:

$$\begin{aligned} a_{\min} &= 23,03, & b_{\min} &= 0,29, \\ a_{\max} &= 130,92, & b_{\max} &= 1,54. \end{aligned}$$

Обе границы для каждого из параметров имеют один знак, следовательно, оба параметра уравнения линейной регрессии существенны.

Расчет прогнозного значения результативной переменной

Прогнозное значение результативной переменной $\hat{y}_p$ рассчитывается по уравнению регрессии при подстановке в него прогнозного значения факторной переменной x_p .

В нашем примере прогнозное значение факторной переменной составляет 107% от среднего уровня

$$x_p = 85,58 \cdot 1,07 = 91,5706.$$

Точечный прогноз среднедневной заработной платы $\hat{y}_p$ при величине среднедушевого прожиточного минимума 91,5706 ден.ед. вычисляем по найденным уравнениям регрессии (табл.8).

Таблица 8

Регрессионная модель	Уравнение регрессии	Прогнозное значение $\hat{y}_p$
линейная	$\hat{y} = 76,976 + 0,9204x$	161,2609
степенная	$\hat{y} = 16,774x^{0,5008}$	161,0957
показательная	$\hat{y} = 95,54 \cdot 1,01^x$	159,5433
полулогарифмическая	$\hat{y} = -206,17 + 81,536 \cdot \ln x$	162,1371
гиперболическая	$\hat{y} = 238,87 - \frac{6966,4}{x}$	162,7932

На основе стандартной ошибки прогноза

$$m_{\hat{y}_p} = S_{\text{ост}} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{n \cdot \sigma_x^2}},$$

рассчитываются границы доверительного интервала

$$\hat{y}_{p_{\min}} = \hat{y}_p - t_{\text{табл}} \cdot m_{\hat{y}_p}, \quad \hat{y}_{p_{\max}} = \hat{y}_p + t_{\text{табл}} \cdot m_{\hat{y}_p}.$$

Относительная величина различий значений верхней и нижней границ прогноза, которая находится по формуле

$$D = \frac{\hat{y}_{p_{\max}}}{\hat{y}_{p_{\min}}},$$

характеризует точность прогноза: $D \leq 2$, $D = 1,2 - 1,3$ – точный прогноз.

Таблица 8

Регрессионная модель	Прогнозное значение $\hat{y}_p$	Верхняя граница $\hat{y}_{p_{\max}}$	Нижняя граница $\hat{y}_{p_{\min}}$	Характеристика точности D
линейная	161,2609	188,0468	134,4749	1,3984
степенная	161,0957	188,0272	134,1705	1,4014
показательная	159,5433	186,3013	132,7853	1,4030
полулогарифмическая	162,1371	189,2538	135,0267	1,4016
гиперболическая	162,7932	190,5005	135,0917	1,4102

Как видно из таблицы, относительная величина различий значений верхней и нижней границ составляет в среднем 1,403. Это означает, что выполненный на уровне 95% надёжности прогноз недостаточно точный. Причиной некоторой неточности прогноза является невысокое значение индекса корреляции у всех моделей, около 70%. Скорее всего, это имеет место из-за неравномерности расположения точек на поле корреляции. Возможно, если дополнить данные так, чтобы количество точек слева и справа от среднего было примерно одинаково, то качество моделей и точность прогноза по ним заметно повысятся.

Сравнительная оценка силы связи фактора с результативной переменной

Расчёт средних коэффициентов эластичности по найденным уравнениям регрессии (табл.9):

Таблица 9

Вид функции	Уравнение	Средний коэффициент
-------------	-----------	---------------------

	регрессии	эластичности, $\bar{\varepsilon}$
линейная	$\hat{y} = 76,976 + 0,9204x$	0,5158
степенная	$\hat{y} = 16,774x^{0,5008}$	0,5008
показательная	$\hat{y} = 95,54 \cdot 1,01^x$	0,5128
полулогарифмическая	$\hat{y} = -206,17 + 81,536 \cdot \ln x$	0,5206
гиперболическая	$\hat{y} = 238,87 - \frac{6966,4}{x}$	0,5169

Из таблицы видно, что при изменении среднедушевого прожиточного минимума на 1% от своего среднего значения величина среднедневной заработной платы изменится в среднем на 0,51%.

Сравнительная оценка качества регрессионных моделей

Качество регрессионной модели в целом можно оценить с помощью критерия Фишера. Наблюдаемое значение F -критерия сравнивается с табличным $F_{\text{табл}} = F_{\alpha; k_1; k_2}$. Если $F_{\text{набл}} > F_{\text{табл}}$, то уравнение регрессии признается статистически значимым и надежным.

В рассматриваемом примере уровень значимости $\alpha = 0,05$, число степеней свободы $k_2 = 12 - 1 - 1 = 10$ (для остаточной суммы квадратов) и $k_1 = 1$ (для факторной суммы квадратов). Табличное значение критерия Фишера определяется с помощью функции ФРАСПОБР (рис.49).

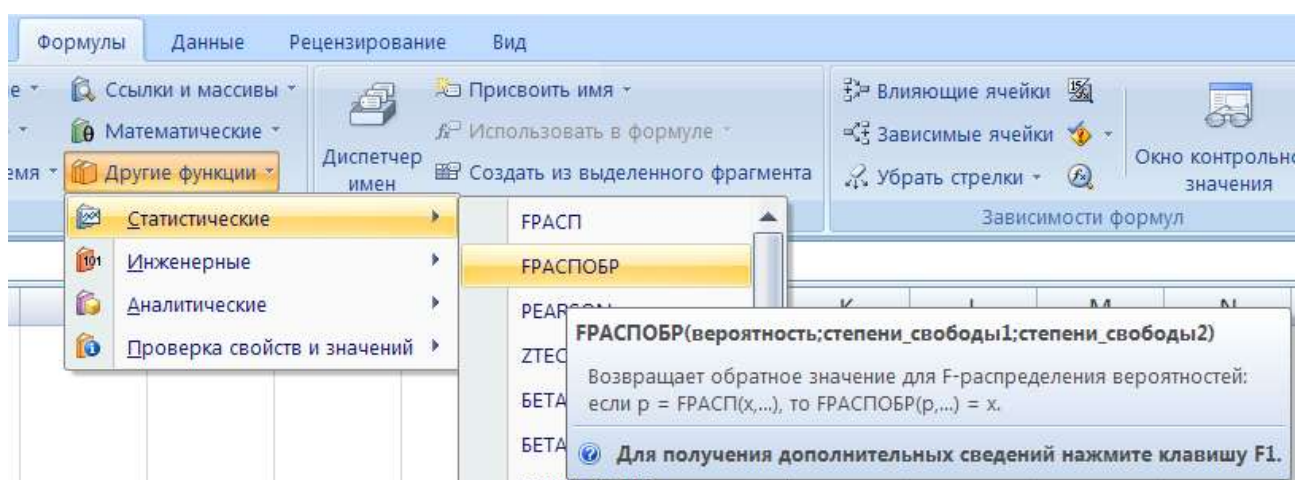


Рис. 49. Выбор функции ФРАСПОБР

В открывшемся окне ввода аргументов указываем наши значения (рис.50) и получаем табличное значение критерия Фишера $F_{\text{табл}} = 4,96$.

Рис. 50. Ввод аргументов функции ФРАСПОБР

Теперь можно провести сравнительный анализ построенных моделей регрессии по значениям рассчитанных характеристик (табл.10).

Поскольку наблюдаемое значение критерия Фишера для всех моделей превышает табличное, то все найденные регрессии можно признать статистически значимыми. Из нелинейных моделей самое маленькое значение $\bar{A}$ имеет степенная регрессия. Самое большое значение ρ_{xy} имеет показательная регрессия. Но наиболее хорошо исходные данные приближает линейная модель. Она имеет наименьшую ошибку аппроксимации и наибольшее значение индекса корреляции. Таким образом, для построения прогноза лучше всего подойдёт модель парной линейной регрессии. Хотя величина коэффициента детерминации $r_{xy}^2 = 0,52$ говорит о том, что в среднем вариация среднедневной заработной платы объясняется вариацией среднедушевого прожиточного минимума только на 52%. Следовательно, данную модель в дальнейшем необходимо дополнить, т.е. ввести в неё дополнительные факторные переменные с тем, чтобы увеличить значение коэффициента детерминации.

Таблица 10

Вид функции	Уравнение регрессии	Средняя ошибка $\bar{A}, \%$	Индекс корреляции ρ_{xy}	F-критерий Фишера
линейная	$\hat{y} = 76,976 + 0,9204x$	5,75	0,721	10,83
степенная	$\hat{y} = 16,774x^{0,5008}$	5,78	0,717	10,58
показательная	$\hat{y} = 95,54 \cdot 1,01^x$	5,89	0,719	10,69

полулогарифмическая	$\hat{y} = -206,17 + 81,536 \cdot \ln x$	5,86	0,713	10,33
гиперболическая	$\hat{y} = 238,87 - \frac{6966,4}{x}$	5,94	0,697	9,47

Варианты индивидуальных заданий

По регионам некоторой территории приводятся данные за год.

Требуется:

1. Постройте поле корреляции (диаграмму рассеяния) и сформулируйте гипотезу о форме связи исследуемых величин.
2. Для характеристики зависимости y от x рассчитайте параметры следующих функций:
 - е) линейной $\hat{y} = a + bx$;
 - ж) степенной $\hat{y} = ax^b$;
 - з) показательной $\hat{y} = ab^x$;
 - и) полулогарифмической $\hat{y} = a + b \cdot \ln x$;
 - к) гиперболической $\hat{y} = a + \frac{b}{x}$.
3. Выполните прогноз среднедневной заработной платы y при прогнозном значении среднедушевого прожиточного минимума x , составляющем 107% от среднего уровня.
4. Оцените тесноту связи с помощью индексов корреляции ρ_{xy} и детерминации ρ_{xy}^2 .
5. Дайте с помощью среднего коэффициента эластичности $\bar{\epsilon}$ сравнительную оценку силы связи фактора с результатом.
6. Оцените качество уравнений регрессии с помощью средней ошибки аппроксимации $\bar{A}$.
7. Оцените с помощью F -критерия Фишера статистическую надежность результатов регрессионного моделирования.
8. Результаты расчетов представьте в табличной форме.

Вариант 1

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного (ден.ед.), x	Среднедневная заработная плата (ден.ед.), y
1	81	124
2	77	131
3	85	146
4	79	139
5	93	143
6	100	159
7	72	135
8	90	152
9	71	127
10	89	154
11	82	127
12	111	162

Вариант 2

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного (ден.ед.), x	Среднедневная заработная плата (ден.ед.), y
1	74	122
2	81	134
3	90	136
4	79	125
5	89	120
6	87	127
7	77	125
8	93	148
9	70	122
10	93	157
11	87	144
12	121	165
13	115	162

Вариант 3

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного, руб., x	Среднедневная заработная плата, руб., y
1	77	123
2	85	152
3	79	140
4	93	142
5	89	157
6	81	141
7	79	133
8	97	163
9	73	134
10	95	155
11	84	132
12	108	165
13	102	151
14	111	165

Вариант 4

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного, руб., х	Среднедневная заработная плата, руб., у
1	83	137
2	88	142
3	75	128
4	89	140
5	85	133
6	79	153
7	81	142
8	97	154
9	79	132
10	90	150
11	84	132
12	112	166
13	81	120
14	114	162
15	105	168

Вариант 5

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного, руб., х	Среднедневная заработная плата, руб., у
1	79	134
2	91	154
3	77	128
4	87	138
5	84	133
6	76	144
7	84	160
8	94	149
9	79	125
10	98	163
11	81	120
12	115	162
13	85	131
14	113	165
15	82	119
16	115	163

Вариант 6

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного, руб., х	Среднедневная заработная плата, руб., у
1	92	147
2	78	133
3	79	128
4	88	152
5	87	138
6	75	122
7	81	145
8	96	141
9	80	127
10	102	151
11	83	129
12	94	147
13	113	156
14	115	162
15	89	150
16	86	129
17	110	159

Вариант 7

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного, руб., х	Среднедневная заработная плата, руб., у
1	75	133
2	78	125
3	81	129
4	93	153
5	86	140
6	77	135
7	83	141
8	94	152
9	88	133
10	99	156
11	80	124
12	112	156
13	113	160
14	114	162
15	100	154

Вариант 8

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного, руб., х	Среднедневная заработная плата, руб., у
1	69	124
2	83	133
3	92	146
4	97	153
5	88	138
6	93	159
7	74	145
8	79	152
9	105	168
10	99	154
11	85	127
12	94	155
13	81	121
14	115	169

Вариант 9

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного, руб., х	Среднедневная заработная плата, руб., у
1	78	133
2	94	139
3	85	141
4	73	127
5	91	154
6	88	142
7	73	122
8	82	135
9	99	142
10	113	168
11	69	124
12	83	130
13	86	140
14	77	135
15	83	141
16	87	144
17	121	165
18	116	159

Вариант 10

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного, руб., х	Среднедневная заработная плата, руб., у
1	97	161
2	73	131
3	79	135
4	99	147
5	86	139
6	91	151
7	85	135
8	77	132
9	89	161
10	95	159
11	72	120
12	115	160
13	105	168
14	99	154
15	80	124
16	112	156
17	113	160
18	115	167
19	85	131
20	113	164

Вариант 11

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного (ден.ед.), х	Среднедневная заработная плата (ден.ед.), у
1	81	124
2	77	131
3	85	146
4	79	139
5	93	143
6	100	159
7	72	135
8	90	152
9	71	127
10	89	154
11	82	127
12	111	162

Вариант 12

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного (ден.ед.), x	Среднедневная заработная плата (ден.ед.), y
1	74	122
2	81	134
3	90	136
4	79	125
5	89	120
6	87	127
7	77	125
8	93	148
9	70	122
10	93	157
11	87	144
12	121	165
13	115	162

Вариант 13

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного, руб., x	Среднедневная заработная плата, руб., y
1	77	123
2	85	152
3	79	140
4	93	142
5	89	157
6	81	141
7	79	133
8	97	163
9	73	134
10	95	155
11	84	132
12	108	165
13	102	151
14	111	165

Вариант 14

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного, руб., х	Среднедневная заработная плата, руб., у
1	83	137
2	88	142
3	75	128
4	89	140
5	85	133
6	79	153
7	81	142
8	97	154
9	79	132
10	90	150
11	84	132
12	112	166
13	81	120
14	114	162
15	105	168

Вариант 15

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного, руб., х	Среднедневная заработная плата, руб., у
1	79	134
2	91	154
3	77	128
4	87	138
5	84	133
6	76	144
7	84	160
8	94	149
9	79	125
10	98	163
11	81	120
12	115	162
13	85	131
14	113	165
15	82	119
16	115	163

Вариант 16

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного, руб., х	Среднедневная заработная плата, руб., у
1	92	147
2	78	133
3	79	128
4	88	152
5	87	138
6	75	122
7	81	145
8	96	141
9	80	127
10	102	151
11	83	129
12	94	147
13	113	156
14	115	162
15	89	150
16	86	129
17	110	159

Вариант 17

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного, руб., х	Среднедневная заработная плата, руб., у
1	75	133
2	78	125
3	81	129
4	93	153
5	86	140
6	77	135
7	83	141
8	94	152
9	88	133
10	99	156
11	80	124
12	112	156
13	113	160
14	114	162
15	100	154

Вариант 18

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного, руб., х	Среднедневная заработная плата, руб., у
1	69	124
2	83	133
3	92	146
4	97	153
5	88	138
6	93	159
7	74	145
8	79	152
9	105	168
10	99	154
11	85	127
12	94	155
13	81	121
14	115	169

Вариант 19

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного, руб., х	Среднедневная заработная плата, руб., у
1	78	133
2	94	139
3	85	141
4	73	127
5	91	154
6	88	142
7	73	122
8	82	135
9	99	142
10	113	168
11	69	124
12	83	130
13	86	140
14	77	135
15	83	141
16	87	144
17	121	165
18	116	159

Вариант 20

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного, руб., х	Среднедневная заработная плата, руб., у
1	97	161
2	73	131
3	79	135
4	99	147
5	86	139
6	91	151
7	85	135
8	77	132
9	89	161
10	95	159
11	72	120
12	115	160
13	105	168
14	99	154
15	80	124
16	112	156
17	113	160
18	115	167
19	85	131
20	113	164