Министерство науки и высшего образования Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ ГИДРОМЕТЕОРОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ

А.А. Чалганова

Построение множественной регрессии и оценка качества модели с использованием табличного процессора Excel

Учебное пособие по дисциплине «Эконометрика»

Санкт-Петербург
РГГМУ
2022

УДК 33

ББК 65.05
Ч-16

Чалганова, Алла Анатольевна.

Построение множественной регрессии и оценка качества модели с использованием табличного процессора Excel. Учебное пособие по дисциплине «Эконометрика» / А.А. Чалганова. – [Текст : электронный]. – Санкт-Петербург : РГГМУ, 2022. – 90 с.

В учебном пособии «Построение множественной регрессии и оценка качества модели с использованием табличного процессора Excel» особое внимание уделяется реализации алгоритма решения задачи построения множественной линейной регрессии путем обращения к функции «Регрессия» модуля «Анализ данных» (надстройка табличного процессора Excel).

Учебное пособие предназначено для студентов всех форм обучения по направлению подготовки 38.03.01 «Экономика».

Учебное пособие по дисциплине «Эконометрика» для студентов всех форм обучения одобрено на заседании кафедры экономики предприятия природопользования и учетных систем от 15 июня 2022 г., протокол № 12.

Введение

Построение эконометрических моделей обуславливает существенный объем вычислений, что требует использования вычислительной техники и соответствующего программного обеспечения, особенно при большом объеме исходных данных. Удобным инструментом для решения эконометрических задач является табличный процессор Excel. Будучи удобной универсальной вычислительной средой, Excel позволяет реализовать алгоритм решения путем программирования арифметических или логических выражений в ячейках электронной таблицы, либо путем обращения к стандартным функциям и модулям.

Использование современных алгоритмов решения эконометрических задач на практике предполагает знакомство с возможностями использования ЭВМ при обработке массовой статистической информации, моделировании и прогнозировании явлений.

Поэтому основной целью данного пособия является изложение численной методики решения основных задач линейного регрессионного анализа в вычислительной среде табличного процессора Excel.

1. Общие методические указания для решения задач по дисциплине «Эконометрика»

Целью данного пособия является рассмотрение проблемы спецификации модели, которая включает в себя два вопроса: отбор факторов и выбор вида уравнения регрессии. Акцент сделан на отборе факторов при построении множественной регрессии.

Решение задач предполагает наличие предварительной теоретической подготовки по той теме дисциплины, для усвоения которой предлагается определенная задача. Чтобы очертить круг вопросов, которые надо прояснить для решения задачи, необходимо, прежде всего, внимательно разобраться в условии задачи и в вопросах, на которые требуется дать ответы. Это позволит целенаправленно проводить работу с конспектом лекций и эффективнее готовиться к аттестации по предмету, используя теоретические материалы, которые помогут в решении задачи. Поскольку решение задач по дисциплине «Эконометрика» связано с расчетом различных показателей, необходимо хорошо ориентироваться в них, что достигается выписыванием терминов, понятий, названий показателей и формул их расчета, а также систематизацией подобных записей. Использование современных программных средств упрощает выполнение расчетов, сокращает время решения задач, расширяет их практическое применение для планирования и составления прогнозов. Определив необходимые для решения задачи показатели и формулы их расчета, выбрав средства автоматизации расчетов, студенты обычно легко справляются с вычислениями, однако полученные цифры не являются конечной целью решения задачи. Задача исследователя, в роли которого выступает решающий учебную задачу студент, не ограничивается получением какого-то числового значения, хотя это важно и полезно само по себе.

Главная цель исследования – дать оценку и интерпретацию полученным значениям, т.е. разъяснить их смысл, оценить статистическую значимость эконометрической модели и ее параметров, сделать вывод о возможности использования данной модели для анализа и прогнозирования. Поэтому решение задачи по эконометрике должно заканчиваться выводами, полученными в результате корреляционно-регрессионного анализа.

2. Множественная регрессия и корреляция

2.1. Экономические процессы, описываемые с помощью уравнений множественной регрессии

Парная регрессия может дать хороший результат при моделировании, если влиянием других факторов, воздействующих на объект исследования, можно пренебречь. Если же этим влиянием пренебречь нельзя, то следует попытаться выявить влияние других факторов, введя их в модель, т.е. построить уравнение множественной регрессии:

$$y = f\left(x_1,\, x_2,\, ...,\, x_m\right),$$

где y – зависимая переменная (результативный признак), x_i – независимые, или объясняющие, переменные (признаки-факторы).

Множественная регрессия широко используется в решении проблем спроса, доходности акций, при изучении функции издержек производства, в макроэкономических расчетах и целом ряде других вопросов эконометрики. В настоящее время множественная регрессия – один из наиболее распространенных методов в эконометрике. Основная цель множественной регрессии – построить модель с большим числом факторов, определив при этом влияние каждого из них в отдельности, а также совокупное их воздействие на моделируемый показатель.

2.2. Спецификация модели. Отбор факторов при построении уравнения множественной регрессии

Построение уравнения множественной регрессии начинается с решения вопроса о спецификации модели. Он включает в себя два круга вопросов: отбор факторов и выбор вида уравнения регрессии.

Включение в уравнение множественной регрессии того или иного набора факторов связано прежде всего с представлением исследователя о природе взаимосвязи моделируемого показателя с другими экономическими явлениями. Факторы, включаемые во множественную регрессию, должны отвечать следующим требованиям:

1. Они должны быть количественно измеримы. Если необходимо включить в модель качественный фактор, не имеющий количественного измерения, то ему нужно придать количественную определенность.
2. Факторы не должны быть интеркоррелированы и тем более находиться в точной функциональной связи.

Включение в модель факторов с высокой интеркорреляцией, может привести к нежелательным последствиям – система нормальных уравнений может оказаться плохо обусловленной и повлечь за собой неустойчивость и ненадежность оценок коэффициентов регрессии.

Если между факторами существует высокая корреляция, то нельзя определить их изолированное влияние на результативный показатель и параметры уравнения регрессии оказываются неинтерпретируемыми.

Включаемые во множественную регрессию факторы должны объяснить вариацию зависимой переменной. Если строится модель с набором m факторов, то для нее рассчитывается показатель детерминации R^2, который фиксирует долю объясненной вариации результативного признака за счет рассматриваемых в регрессии m факторов. Влияние других, не учтенных в модели факторов, оценивается как $1-R^2$ с соответствующей остаточной дисперсией S^2.

При дополнительном включении в регрессию $m+1$ фактора коэффициент детерминации должен возрастать, а остаточная дисперсия уменьшаться:

$$R_{m+1}^2 \geq R_m^2 \quad \text{и} \qquad S_{m+1}^2 \leq S_m^2.$$

Если же этого не происходит и данные показатели практически не отличаются друг от друга, то включаемый в анализ фактор x_{m+1} не улучшает модель и практически является лишним фактором.

Насыщение модели лишними факторами не только не снижает величину остаточной дисперсии и не увеличивает показатель детерминации, но и приводит к статистической незначимости параметров регрессии по критерию Стьюдента.

Таким образом, хотя теоретически регрессионная модель позволяет учесть любое число факторов, практически в этом нет необходимости. Отбор факторов производится на основе качественного теоретико-экономического анализа. Однако теоретический анализ часто не позволяет однозначно ответить на вопрос о количественной взаимосвязи рассматриваемых признаков и целесообразности включения фактора в модель. Поэтому отбор факторов обычно осуществляется в две стадии: на первой подбираются факторы исходя из сущности проблемы; на второй – на основе матрицы показателей корреляции определяют статистики для параметров регрессии.

Коэффициенты интеркорреляции (т.е. корреляции между объясняющими переменными) позволяют исключать из модели дублирующие факторы. Считается, что две переменные явно коллинеарны, т.е. находятся между собой в линейной зависимости, если $r_{x_i x_j} \geq 0,7$. Если факторы явно коллинеарны, то они дублируют друг друга и один из них рекомендуется исключить из регрессии. Предпочтение при этом отдается не фактору, более тесно связанному с результатом, а тому фактору, который при достаточно тесной связи с результатом имеет наименьшую тесноту связи с другими факторами. В этом требовании проявляется специфика множественной регрессии как метода исследования комплексного воздействия факторов в условиях их независимости друг от друга.

Пусть, например, при изучении зависимости $y = f(x_1, x_2, x_3)$ матрица парных коэффициентов корреляции, приведенных в таблице 2.1, оказалась следующей:

Таблица 2.1

	y	x_1	x_2	x_3
y	1	0,8	0,7	0,6
x_1	0,8	1	0,8	0,5
x_2	0,7	0,8	1	0,2
x_3	0,6	0,5	0,2	1

Очевидно, что факторы x_1 и x_2 дублируют друг друга ($r_{x_1x_2}$ = 0,8). В анализ целесообразно включить фактор x_2, а не x_1, хотя корреляция x_2 с результатом y слабее, чем корреляция фактора x_1 с y $\left(r_{yx_2} = 0,7 < r_{yx_1} = 0,8\right)$, но зато значительно слабее межфакторная корреляция: $r_{x_2x_3} = 0,2 < r_{x_1x_3} = 0,5$. Поэтому в данном случае в уравнение множественной регрессии включаются факторы x_2, x_3.

По величине парных коэффициентов корреляции обнаруживается лишь явная коллинеарность факторов. Наибольшие трудности в использовании аппарата множественной регрессии возникают при наличии **мультиколлинеарности факторов**, когда более чем два фактора связаны между собой линейной зависимостью, т.е. имеет место совокупное воздействие факторов друг на друга. Наличие мультиколлинеарности факторов может означать, что некоторые факторы будут всегда действовать в унисон. В результате вариация в исходных данных перестает быть полностью независимой и нельзя оценить воздействие каждого фактора в отдельности.

Включение в модель мультиколлинеарных факторов нежелательно в силу следующих последствий:

1. Затрудняется интерпретация параметров множественной регрессии как характеристик действия факторов в «чистом» виде, ибо факторы коррелированы; параметры линейной регрессии теряют экономический смысл.

2. Оценки параметров ненадежны, обнаруживают большие стандартные ошибки и меняются с изменением объема наблюдений (не только по величине, но и по знаку), что делает модель непригодной для анализа и прогнозирования.

Для оценки мультиколлинеарности факторов может использоваться определитель матрицы парных коэффициентов корреляции между факторами.

Если бы факторы не коррелировали между собой, то матрица парных коэффициентов корреляции между факторами была бы единичной матрицей, поскольку все недиагональные элементы $r_{x_i x_j}$ $(i \neq j)$ были бы равны нулю. Так, для уравнения, включающего три объясняющих переменных

$$y = a + b_1 x_1 + b_2 x_2 + b_3 x_3$$

матрица коэффициентов корреляции между факторами имела бы определитель, равный единице:

$$\operatorname{Det} \mathbf{R} = \begin{vmatrix} r_{x_1x_1} & r_{x_1x_2} & r_{x_1x_3} \\ r_{x_2x_1} & r_{x_2x_2} & r_{x_2x_3} \\ r_{x_3x_1} & r_{x_3x_2} & r_{x_3x_3} \end{vmatrix} = \begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix} = 1.$$

Если же, наоборот, между факторами существует полная линейная зависимость и все коэффициенты корреляции равны единице, то определитель такой матрицы равен нулю:

$$\operatorname{Det} \mathbf{R} = \begin{vmatrix} r_{x_1x_1} & r_{x_1x_2} & r_{x_1x_3} \\ r_{x_2x_1} & r_{x_2x_2} & r_{x_2x_3} \\ r_{x_3x_1} & r_{x_3x_2} & r_{x_3x_3} \end{vmatrix} = \begin{vmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{vmatrix} = 0.$$

Чем ближе к нулю определитель матрицы межфакторной корреляции, тем сильнее мультиколлинеарность факторов и ненадежнее результаты множественной регрессии. И наоборот, чем ближе к единице определитель матрицы межфакторной корреляции, тем меньше мультиколлинеарность факторов.

Существует ряд подходов преодоления сильной межфакторной корреляции. Самый простой путь устранения мультиколлинеарности состоит в исключении из модели одного или нескольких факторов. Другой подход связан с преобразованием факторов, при котором уменьшается корреляция между ними.

Одним из путей учета внутренней корреляции факторов является переход к совмещенным уравнениям регрессии, т.е. к уравнениям, которые отражают не только влияние факторов, но и их взаимодействие. Так, если $y = f\left(x_1,\ x_2, x_3\right)$, то возможно построение следующего совмещенного уравнения:

$$y = a + b_1x_1 + b_2x_2 + b_3x_3 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{23}x_2x_3 + \varepsilon .$$

Рассматриваемое уравнение включает взаимодействие первого порядка (взаимодействие двух факторов). Возможно включение в модель и взаимодействий более высокого порядка, если будет доказана их статистическая значимость по F-критерию Фишера, но, как правило, взаимодействия третьего и более высоких порядков оказываются статистически незначимыми.

Отбор факторов, включаемых в регрессию, является одним из важнейших этапов практического использования методов регрессии. Подходы к отбору факторов на основе показателей корреляции могут быть разные. Они приводят построение уравнения множественной регрессии соответственно к разным методикам. В зависимости от того, какая методика построения уравнения регрессии принята, меняется алгоритм ее решения на ЭВМ.

Наиболее широкое применение получили следующие методы построения уравнения множественной регрессии:

1. Метод исключения – отсев факторов из полного его набора.

2. Метод включения – дополнительное введение фактора.

3. Шаговый регрессионный анализ – исключение ранее введенного фактора.

При отборе факторов также рекомендуется пользоваться следующим правилом: число включаемых факторов обычно в 6–7 раз меньше объема совокупности, по которой строится регрессия. Если это соотношение нарушено, то число степеней свободы остаточной дисперсии очень мало. Это приводит к тому, что параметры уравнения регрессии оказываются статистически незначимыми, а F-критерий меньше табличного значения.

2.3. Метод наименьших квадратов (МНК). Свойства оценок на основе МНК

Возможны разные виды уравнений множественной регрессии: линейные и нелинейные.

Ввиду четкой интерпретации параметров наиболее широко используется линейная функция. В линейной множественной регрессии $y_x = a + b_1 x_1 + b_2 x_2 + ... + b_m x_m$ параметры при x называются коэффициентами «чистой» регрессии. Они характеризуют среднее изменение результата с изменением соответствующего фактора на единицу при неизмененном значении других факторов, закрепленных на среднем уровне.

Рассмотрим линейную модель множественной регрессии

$$y = a + b_1 x_1 + b_2 x_2 + ... + b_m x_m + \varepsilon. \tag{2.1}$$

Классический подход к оцениванию параметров линейной модели множественной регрессии основан на методе наименьших квадратов (МНК). МНК позволяет получить такие оценки параметров, при которых сумма квадратов отклонений фактических значений результативного признака y от расчетных y минимальна:

$$\sum_i \left(y_i - y_{x_i} \right)^2 \quad \rightarrow \quad \min. \tag{2.2}$$

Как известно из курса математического анализа, для того чтобы найти экстремум функции нескольких переменных, надо вычислить частные производные первого порядка по каждому из параметров и приравнять их к нулю.

Итак, имеем функцию $m+1$ аргумента:

$$S(a,\ b_1,\ b_2,\ ...,\ b_m) = \sum (y - a - b_1x_1 - b_2x_2 - ... - b_mx_m)^2 .$$

Находим частные производные первого порядка:

$$\begin{cases} \dfrac{\partial S}{\partial a} = -2\sum (y - a - b_1x_1 - b_2x_2 - ... - b_mx_m) = 0; \\ \dfrac{\partial S}{\partial b_1} = -2\sum x_1 (y - a - b_1x_1 - b_2x_2 - ... - b_mx_m) = 0; \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ \dfrac{\partial S}{\partial b_m} = -2\sum x_m (y - a - b_1x_1 - b_2x_2 - ... - b_mx_m) = 0. \end{cases}$$

После элементарных преобразований приходим к системе линейных нормальных уравнений для нахождения параметров линейного уравнения множественной регрессии (2.1):

$$\begin{cases} na + b_1\sum x_1 + b_2\sum x_2 + ... + b_m\sum x_m = \sum y, \\ a\sum x_1 + b_1\sum x_1^2 + b_2\sum x_1x_2 + ... + b_m\sum x_1x_m = \sum yx_1, \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ a\sum x_m + b_1\sum x_1x_m + b_2\sum x_2x_m + ... + b_m\sum x_m^2 = \sum yx_m. \end{cases} \qquad (2.3)$$

Для двухфакторной модели данная система будет иметь вид:

$$\begin{cases} na + b_1\sum x_1 + b_2\sum x_2 = \sum y, \\ a\sum x_1 + b_1\sum x_1^2 + b_2\sum x_1x_2 = \sum yx_1, \\ a\sum x_2 + b_1\sum x_1x_2 + b_2\sum x_2^2 = \sum yx_2. \end{cases}$$

Метод наименьших квадратов применим и к уравнению множественной регрессии в стандартизированном масштабе:

$$t_y = \beta_1 t_{x_1} + \beta_2 t_{x_2} + ... + \beta_m t_{x_m} + \varepsilon, \qquad (2.4)$$

где t_y, t_{x_1}, ..., t_{x_m} – стандартизированные переменные: $t_y = \dfrac{y-\overline{y}}{\sigma_y}$, $t_{x_i} = \dfrac{x_i-\overline{x}_i}{\sigma_{x_i}}$, для которых среднее значение равно нулю: $\overline{t_y} = \overline{t_{x_i}} = 0$, а среднее квадратическое отклонение равно единице: $\sigma_{t_y} = \sigma_{t_{x_i}} = 1$; β_i – стандартизированные коэффициенты регрессии.

Стандартизованные коэффициенты регрессии показывают, на сколько единиц изменится в среднем результат, если соответствующий фактор x_i изменится на одну единицу при неизменном среднем уровне других факторов. В силу того, что все переменные заданы как центрированные и нормированные, стандартизованные коэффициенты регрессии β_i можно сравнивать между собой. Сравнивая их друг с другом, можно ранжировать факторы по силе их воздействия на результат. В этом основное достоинство стандартизованных коэффициентов регрессии в отличие от коэффициентов «чистой» регрессии, которые несравнимы между собой.

Применяя МНК к уравнению множественной регрессии в стандартизированном масштабе, получим систему нормальных уравнений вида

$$\begin{cases} r_{yx_1} = \beta_1 + \beta_2 r_{x_1x_2} + \beta_3 r_{x_1x_3} + \ldots + \beta_m r_{x_1x_m}, \\ r_{yx_2} = \beta_1 r_{x_1x_2} + \beta_2 + \beta_3 r_{x_1x_3} + \ldots + \beta_m r_{x_1x_m}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ r_{yx_m} = \beta_1 r_{x_1x_m} + \beta_2 r_{x_2x_m} + \beta_3 r_{x_3x_m} + \ldots + \beta_m, \end{cases} \tag{2.5}$$

где r_{yx_i} и $r_{x_ix_j}$ – коэффициенты парной и межфакторной корреляции.

Коэффициенты «чистой» регрессии b_i связаны со стандартизованными коэффициентами регрессии β_i следующим образом:

$$b_i = \beta_i \frac{\sigma_y}{\sigma_{x_i}}. \tag{2.6}$$

Поэтому можно переходить от уравнения регрессии в стандартизованном масштабе (2.4) к уравнению регрессии в натуральном масштабе переменных (2.1), при этом параметр a определяется как $a = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 - \ldots - b_m\bar{x}_m$.

Рассмотренный смысл стандартизованных коэффициентов регрессии позволяет их использовать при отсеве факторов – из модели исключаются факторы с наименьшим значением β_i.

На основе линейного уравнения множественной регрессии

$$y = a + b_1x_1 + b_2x_2 + \ldots + b_mx_m + \varepsilon \tag{2.7}$$

могут быть найдены частные уравнения регрессии:

$$\begin{cases} y_{x_1\cdot x_2, x_3, \ldots, x_m} = f\left(x_1\right), \\ y_{x_2\cdot x_1, x_3, \ldots, x_m} = f\left(x_2\right), \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_{x_m\cdot x_1, x_2, \ldots, x_{m-1}} = f\left(x_m\right), \end{cases} \tag{2.8}$$

т.е. уравнения регрессии, которые связывают результативный признак с соответствующим фактором x_i при закреплении остальных факторов на среднем уровне. В развернутом виде систему (3.8) можно переписать в виде:

$$\begin{cases} y_{x_1\cdot x_2, x_3, \ldots, x_m} = a + b_1x_1 + b_2\bar{x}_2 + b_3\bar{x}_3 + \ldots + b_m\bar{x}_m + \varepsilon, \\ y_{x_2\cdot x_1, x_3, \ldots, x_m} = a + b_1\bar{x}_1 + b_2x_2 + b_3\bar{x}_3 + \ldots + b_m\bar{x}_m + \varepsilon, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_{x_m\cdot x_1, x_2, \ldots, x_{m-1}} = a + b_1\bar{x}_1 + b_2\bar{x}_2 + b_3\bar{x}_3 + \ldots + b_mx_m + \varepsilon. \end{cases}$$

При подстановке в эти уравнения средних значений соответствующих факторов они принимают вид парных уравнений линейной регрессии, т.е. имеем

$$
\begin{cases}
y_{x_1 \cdot x_2, x_3, \ldots, x_m} = A_1 + b_1 x_1, \\
y_{x_2 \cdot x_1, x_3, \ldots, x_m} = A_2 + b_2 x_2, \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
y_{x_m \cdot x_1, x_2, \ldots, x_{m-1}} = A_m + b_m x_m,
\end{cases} \tag{2.9}
$$

где

$$
\begin{cases}
A_1 = a + b_2 \bar{x}_2 + b_3 \bar{x}_3 + \ldots + b_m \bar{x}_m, \\
A_2 = a + b_1 \bar{x}_1 + b_3 \bar{x}_3 + \ldots + b_m \bar{x}_m, \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
A_m = a + b_1 \bar{x}_1 + b_2 \bar{x}_2 + b_3 \bar{x}_3 + \ldots + b_{m-1} x_{m-1}.
\end{cases}
$$

В отличие от парной регрессии частные уравнения регрессии характеризуют изолированное влияние фактора на результат, ибо другие факторы закреплены на неизменном уровне. Эффекты влияния других факторов присоединены в них к свободному члену уравнения множественной регрессии. Это позволяет на основе частных уравнений регрессии определять частные коэффициенты эластичности:

$$
Э_{y_{x_i}} = b_i \cdot \frac{x_i}{y_{x_i \cdot x_1, x_2, \ldots x_{i-1}, x_{i+1}, \ldots, x_m}}, \tag{2.10}
$$

где b_i – коэффициент регрессии для фактора x_i в уравнении множественной регрессии, $y_{x_i \cdot x_1, x_2, \ldots x_{i-1}, x_{i+1}, \ldots, x_m}$ – частное уравнение регрессии.

Наряду с частными коэффициентами эластичности могут быть найдены средние по совокупности показатели эластичности:

$$
\bar{Э}_i = b_i \cdot \frac{\bar{x}_i}{\bar{y}_{x_i}}, \tag{2.11}
$$

Средние показатели эластичности показывают на сколько процентов в среднем изменится результат при изменении соответствующего фактора на 1%. Их можно сравнивать друг с другом и соответственно ранжировать факторы по силе воздействия на результат.

2.4. Проверка существенности факторов и показатели качества регрессии

Практическая значимость уравнения множественной регрессии оценивается с помощью показателя множественной корреляции и его квадрата – показателя детерминации.

Показатель множественной корреляции характеризует тесноту связи рассматриваемого набора факторов с исследуемым признаком или, иначе, оценивает тесноту совместного влияния факторов на результат.

Независимо от формы связи показатель множественной корреляции может быть найден как индекс множественной корреляции:

$$R_{yx_1x_2\ldots x_m} = \sqrt{1 - \frac{\sigma^2_{\text{ост}}}{\sigma^2_y}}, \tag{2.12}$$

где σ^2_y – общая дисперсия результативного признака;

$\sigma^2_{\text{ост}}$ – остаточная дисперсия.

Границы изменения индекса множественной корреляции от 0 до 1. Чем ближе его значение к 1, тем теснее связь результативного признака со всем набором исследуемых факторов. Величина индекса множественной корреляции должна быть больше или равна максимальному парному индексу корреляции:

$$R_{yx_1x_2\ldots x_m} \geq r_{yx_i\,(\max)} \quad \left(i = \overline{1,m}\right).$$

При правильном включении факторов в регрессионную модель величина индекса множественной корреляции будет существенно отличаться от индекса корреляции парной зависимости. Если же дополнительно включенные в уравнение множественной регрессии факторы третьестепенны, то индекс множественной корреляции может практически совпадать с индексом парной корреляции (различия в третьем, четвертом знаках). Отсюда ясно, что, сравнивая индексы множественной и парной корреляции, можно сделать вывод о целесообразности включения в уравнение регрессии того или иного фактора.

Расчет индекса множественной корреляции предполагает определение уравнения множественной регрессии и на его основе остаточной дисперсии:

$$\sigma^2_{\text{ост}} = \frac{1}{n}\sum\left(y - y_{x_1x_2\ldots x_m}\right)^2. \qquad (2.13)$$

Можно пользоваться следующей формулой индекса множественной детерминации:

$$R^2_{yx_1x_2\ldots x_m} = 1 - \frac{\sum\left(y - y_{x_1x_2\ldots x_m}\right)^2}{\sum\left(y - \bar{y}\right)^2}. \qquad (2.14)$$

При линейной зависимости признаков формула индекса множественной корреляции может быть представлена следующим выражением:

$$R_{yx_1x_2\ldots x_m} = \sqrt{\sum \beta_i \cdot r_{yx_i}}, \qquad (2.15)$$

где β_i – стандартизованные коэффициенты регрессии;

r_{yx_i} – парные коэффициенты корреляции результата с каждым фактором.

Формула индекса множественной корреляции для линейной регрессии получила название *линейного коэффициента множественной корреляции*, или, что то же самое, *совокупного коэффициента корреляции*.

Возможно также при линейной зависимости определение совокупного коэффициента корреляции через матрицу парных коэффициентов корреляции:

$$R_{yx_1x_2,\ldots,x_m} = \sqrt{1 - \frac{\Delta r}{\Delta r_{11}}},$$

(2.16)

где

$$\Delta r = \begin{vmatrix} 1 & r_{yx_1} & r_{yx_2} & \dots & r_{yx_m} \\ r_{yx_1} & 1 & r_{x_1x_2} & \dots & r_{x_1x_m} \\ r_{yx_2} & r_{x_2x_1} & 1 & \dots & r_{x_2x_m} \\ \dots & \dots & \dots & \dots & \dots \\ r_{yx_m} & r_{x_mx_1} & r_{x_mx_2} & \dots & 1 \end{vmatrix}$$

– определитель матрицы парных коэффициентов корреляции;

$$\Delta r_{11} = \begin{vmatrix} 1 & r_{x_1x_2} & \dots & r_{x_1x_m} \\ r_{x_2x_1} & 1 & \dots & r_{x_2x_m} \\ \dots & \dots & \dots & \dots \\ r_{x_mx_1} & r_{x_mx_2} & \dots & 1 \end{vmatrix}$$

– определитель матрицы межфакторной корреляции.

Как видим, величина множественного коэффициента корреляции зависит не только от корреляции результата с каждым из факторов, но и от межфакторной корреляции. Рассмотренная формула позволяет определять совокупный коэффициент корреляции, не обращаясь при этом к уравнению множественной регрессии, а используя лишь парные коэффициенты корреляции.

В рассмотренных показателях множественной корреляции (индекс и коэффициент) используется остаточная дисперсия, которая имеет систематическую ошибку в сторону преуменьшения, тем более значительную, чем больше параметров определяется в уравнении регрессии при заданном объеме наблюдений n. Если число параметров при x_i равно m и приближается к объему наблюдений, то остаточная дисперсия будет близка к нулю и коэффициент (индекс) корреляции приблизится к единице даже при слабой связи факторов с результатом. Для того чтобы не допустить возможного преувеличения тесноты связи, используется скорректированный индекс (коэффициент) множественной корреляции.

Скорректированный индекс множественной корреляции содержит поправку на число степеней свободы, а именно, остаточная сумма квадратов $\sum\left(y - y_{x_1 x_2 \ldots x_m}\right)^2$ делится на число степеней свободы остаточной вариации $\left(n - m - 1\right)$, а общая сумма квадратов отклонений $\sum\left(y - \bar{y}\right)^2$ на число степеней свободы в целом по совокупности $\left(n - 1\right)$.

Формула скорректированного индекса множественной детерминации имеет вид:

$$R^2 = 1 - \frac{\sum\left(y - y\right)^2 \Big/ \left(n - m - 1\right)}{\sum\left(y - \bar{y}\right) \Big/ \left(n - 1\right)}, \qquad (2.17)$$

где m – число параметров при переменных x; n – число наблюдений.

Поскольку $\dfrac{\sum\left(y - y_{x_1 x_2 \ldots x_m}\right)^2}{\sum\left(y - \bar{y}\right)^2} = 1 - R^2$, то величину скорректированного индекса детерминации можно представить в виде:

$$R^2 = 1 - \left(1 - R^2\right) \cdot \frac{n - 1}{n - m - 1}. \qquad (2.17а)$$

Чем больше величина m, тем сильнее различия R^2 и R^2.

Как было показано выше, ранжирование факторов, участвующих во множественной линейной регрессии, может быть проведено через стандартизованные коэффициенты регрессии (β-коэффициенты). Эта же цель может быть достигнута с помощью частных коэффициентов корреляции (для линейных связей). Кроме того, частные показатели корреляции широко используются при решении проблемы отбора факторов: целесообразность включения того или иного фактора в модель можно доказать величиной показателя частной корреляции.

Частные коэффициенты корреляции характеризуют тесноту связи между результатом и соответствующим фактором при элиминировании (устранении влияния) других факторов, включенных в уравнение регрессии.

Показатели частной корреляции представляют собой отношение сокращения остаточной дисперсии за счет дополнительного включения в анализ нового фактора к остаточной дисперсии, имевшей место до введения его в модель.

В общем виде при наличии m факторов для уравнения

$$y = a + b_1 x_1 + b_2 x_2 ... + b_m x_m + \varepsilon$$

коэффициент частной корреляции, измеряющий влияние на y фактора x_i, при неизменном уровне других факторов, можно определить по формуле:

$$r_{yx_i \cdot x_1 x_2 ... x_{i-1} x_{i+1} ... x_m} = \sqrt{1 - \frac{1 - R^2_{yx_1 x_2 ... x_i ... x_m}}{1 - R^2_{yx_1 x_2 ... x_{i-1} x_{i+1} ... x_m}}}, \qquad (2.18)$$

где $R^2_{yx_1 x_2 ... x_i ... x_m}$ – множественный коэффициент детерминации всех m факторов с результатом;

$R^2_{yx_1 x_2 ... x_{i-1} x_{i+1} ... x_m}$ – тот же показатель детерминации, но без введения в модель фактора x_i.

При двух факторах формула (2.18) примет вид:

$$r_{yx_1 \cdot x_2} = \sqrt{1 - \frac{1 - R^2_{yx_1 x_2}}{1 - r^2_{yx_2}}}; \qquad r_{yx_2 \cdot x_1} = \sqrt{1 - \frac{1 - R^2_{yx_1 x_2}}{1 - r^2_{yx_1}}}. \qquad (2.18а)$$

Порядок частного коэффициента корреляции определяется количеством факторов, влияние которых исключается. Например, $r_{yx_1 \cdot x_2}$ – коэффициент частной корреляции первого порядка. Соответственно коэффициенты парной корреляции называются коэффициентами нулевого порядка. Коэффициенты частной корреляции более высоких порядков можно определить через коэффициенты частной корреляции более низких порядков по рекуррентной формуле:

$$r_{yx_i \cdot x_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_m} = \frac{r_{yx_i \cdot x_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_{m-1}} - r_{yx_m \cdot x_1 x_2 \ldots x_{m-1}} \cdot r_{x_i x_m \cdot x_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_{m-1}}}{\sqrt{\left(1 - r^2_{yx_m \cdot x_1 x_2 \ldots x_{m-1}}\right) \cdot \left(1 - r^2_{x_i x_m \cdot x_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_{m-1}}\right)}}. \quad (2.19)$$

При двух факторах данная формула примет вид:

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1 x_2}}{\sqrt{\left(1 - r^2_{yx_2}\right) \cdot \left(1 - r^2_{x_1 x_2}\right)}}; \; r_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1 x_2}}{\sqrt{\left(1 - r^2_{yx_1}\right) \cdot \left(1 - r^2_{x_1 x_2}\right)}}. \quad (2.19а)$$

Для уравнения регрессии с тремя факторами частные коэффициенты корреляции второго порядка определяются на основе частных коэффициентов корреляции первого порядка. Так, по уравнению $y = a + b_1 x_1 + b_2 x_2 + b_3 x_3 + \varepsilon$ возможно исчисление трех частных коэффициентов корреляции второго порядка:

$$r_{yx_1 \cdot x_2 x_3}, \; r_{yx_2 \cdot x_1 x_3}, \; r_{yx_3 \cdot x_1 x_2},$$

каждый из которых определяется по рекуррентной формуле. Например, при $i = 1$ имеем формулу для расчета $r_{yx_1 \cdot x_2 x_3}$:

$$r_{yx_1 \cdot x_2 x_3} = \frac{r_{yx_1 \cdot x_2} - r_{yx_3 \cdot x_1 x_2} \cdot r_{x_1 x_3 \cdot x_2}}{\sqrt{\left(1 - r^2_{yx_3 \cdot x_1 x_2}\right)\left(1 - r^2_{x_1 x_3 \cdot x_2}\right)}}. \quad (2.20)$$

Рассчитанные по рекуррентной формуле частные коэффициенты корреляции изменяются в пределах от –1 до +1, а по формулам через множественные коэффициенты детерминации – от 0 до 1. Сравнение их друг с другом позволяет ранжировать факторы по тесноте их связи с результатом. Частные коэффициенты корреляции дают меру тесноты связи каждого фактора с результатом в чистом виде. Если из стандартизованного уравнения регрессии $t_y = \beta_1 t_{x_1} + \beta_2 t_{x_2} + \beta_3 t_{x_3} + \varepsilon$ следует, что $\beta_1 > \beta_2 > \beta_3$, т.е. по силе влияния на результат порядок факторов таков: x_1, x_2, x_3, то этот же порядок факторов определяется и по соотношению частных коэффициентов корреляции, $r_{yx_1 \cdot x_2 x_3} > r_{yx_2 \cdot x_1 x_3} > r_{yx_3 \cdot x_1 x_2}$.

В эконометрике частные коэффициенты корреляции обычно не имеют самостоятельного значения. Их используют на стадии формирования модели. Так, при построении многофакторной модели, на первом шаге определяется уравнение регрессии с полным набором факторов и рассчитывается матрица частных коэффициентов корреляции. На втором шаге отбирается фактор с наименьшей и несущественной по t-критерию Стьюдента величиной показателя частной корреляции. Исключив его из модели, строят новое уравнение регрессии. Процедура продолжается до тех пор, пока не окажется, что все частные коэффициенты корреляции существенно отличаются от нуля. Если исключен несущественный фактор, то множественные коэффициенты детерминации на двух смежных шагах построения регрессионной модели почти не отличаются друг от друга, $R^2_{m+1} \approx R^2_m$, где m – число факторов.

Из приведенных выше формул частных коэффициентов корреляции видна связь этих показателей с совокупным коэффициентом корреляции. Зная частные коэффициенты корреляции (последовательно первого, второго и более высокого порядка), можно определить совокупный коэффициент корреляции по формуле:

$$R_{yx_1 x_2 \ldots x_m} = \sqrt{1 - \left(1 - r^2_{yx_1}\right) \cdot \left(1 - r^2_{yx_2 \cdot x_1}\right) \cdot \left(1 - r^2_{yx_3 \cdot x_1 x_2}\right) \cdot \ldots \cdot \left(1 - r^2_{yx_m \cdot x_1 x_2 \ldots x_{m-1}}\right)}. \quad (2.21)$$

В частности, для двухфакторного уравнения формула (2.21) принимает вид:

$$R_{yx_1 x_2 \ldots x_m} = \sqrt{1 - \left(1 - r^2_{yx_1}\right) \cdot \left(1 - r^2_{yx_2 \cdot x_1}\right)}. \quad (2.21)$$

При полной зависимости результативного признака от исследуемых факторов коэффициент совокупного их влияния равен единице. Из единицы вычитается доля остаточной вариации результативного признака $\left(1-r^2\right)$, обусловленная последовательно включенными в анализ факторами. В результате подкоренное выражение характеризует совокупное действие всех исследуемых факторов.

Значимость уравнения множественной регрессии в целом, так же, как и в парной регрессии, оценивается с помощью F-критерия Фишера:

$$F = \frac{S_{\text{факт}}}{S_{\text{ост}}} = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m}, \tag{2.22}$$

где $S_{\text{факт}}$ – факторная сумма квадратов на одну степень свободы;

$S_{\text{ост}}$ – остаточная сумма квадратов на одну степень свободы;

R^2 – коэффициент (индекс) множественной детерминации;

m – число параметров при переменных x (в линейной регрессии совпадает с числом включенных в модель факторов);

n – число наблюдений.

Должна оцениваться значимость не только уравнения в целом, но и фактора, дополнительно включенного в регрессионную модель.

Необходимость такой оценки связана с тем, что не каждый фактор, вошедший в модель, может существенно увеличивать долю объясненной вариации результативного признака. Кроме того, при наличии в модели нескольких факторов они могут вводиться в модель в разной последовательности. Ввиду корреляции между факторами значимость одного и того же фактора может быть разной в зависимости от последовательности его введения в модель.

Мерой для оценки включения фактора в модель служит частный F-критерий, т.е. F_{x_i}.

Частный F-критерий построен на сравнении прироста факторной дисперсии, обусловленного влиянием дополнительно включенного фактора, с остаточной дисперсией на одну степень свободы по регрессионной модели в целом.

В общем виде для фактора x_i частный F-критерий определится как

$$F_{x_i} = \frac{R^2_{yx_1 \ldots x_i \ldots x_m} - R^2_{yx_1 \ldots x_{i-1} x_{i+1} \ldots x_m}}{1 - R^2_{yx_1 \ldots x_i \ldots x_m}} \cdot \frac{n - m - 1}{1}, \tag{2.23}$$

где $R^2_{yx_1 \ldots x_i \ldots x_m}$ – коэффициент множественной детерминации для модели с полным набором факторов,

$R^2_{yx_1 \ldots x_{i-1} x_{i+1} \ldots x_m}$ – тот же показатель, но без включения в модель фактора x_i,

n – число наблюдений,

m – число параметров в модели (без свободного члена).

Фактическое значение частного F-критерия сравнивается с табличным при уровне значимости α и числе степеней свободы: 1 и $n - m - 1$. Если фактическое значение F_{x_i} превышает $F_{\text{табл}}(\alpha,\ k_1,\ k_2)$, то дополнительное включение фактора x_i в модель статистически оправданно и коэффициент чистой регрессии b_i при факторе x_i статистически значим. Если же фактическое значение F_{x_i} меньше табличного, то дополнительное включение в модель фактора x_i не увеличивает существенно долю объясненной вариации признака y, следовательно, нецелесообразно его включение в модель; коэффициент регрессии при данном факторе в этом случае статистически незначим.

Для двухфакторного уравнения частные F-критерии имеют вид:

$$F_{x_1} = \frac{R^2_{yx_1x_2} - r^2_{yx_2}}{1 - R^2_{yx_1x_2}} \cdot (n - 3), \quad F_{x_2} = \frac{R^2_{yx_1x_2} - r^2_{yx_1}}{1 - R^2_{yx_1x_2}} \cdot (n - 3). \tag{2.23а}$$

С помощью частного F-критерия можно проверить значимость всех коэффициентов регрессии в предположении, что каждый соответствующий фактор x_i вводился в уравнение множественной регрессии последним.

Частный F-критерий оценивает значимость коэффициентов чистой регрессии. Зная величину F_{x_i}, можно определить и t-критерий для коэффициента регрессии при i-м факторе, t_{b_i}, а именно:

$$t_{b_i} = \sqrt{F_{x_i}}\,. \tag{2.24}$$

Оценка значимости коэффициентов чистой регрессии по t-критерию Стьюдента может быть проведена и без расчета частных F-критериев. В этом случае, как и в парной регрессии, для каждого фактора используется формула:

$$t_{b_i} = \frac{b_i}{m_{b_i}}, \tag{2.25}$$

где b_i – коэффициент чистой регрессии при факторе x_i, m_{b_i} – средняя квадратическая (стандартная) ошибка коэффициента регрессии b_i.

Для уравнения множественной регрессии $y = a + b_1x_1 + b_2x_2 + \ldots + b_mx_m$ средняя квадратическая ошибка коэффициента регрессии может быть определена по следующей формуле:

$$m_{b_i} = \frac{\sigma_y\sqrt{1-R^2_{yx_1\ldots x_m}}}{\sigma_{x_i}\sqrt{1-R^2_{x_ix_1\ldots x_m}}}\cdot\frac{1}{\sqrt{n-m-1}}, \tag{2.26}$$

где σ_y – среднее квадратическое отклонение для признака y,

σ_{x_i} – среднее квадратическое отклонение для признака x_i,

$R^2_{yx_1\ldots x_m}$ – коэффициент детерминации для уравнения множественной регрессии,

$R^2_{x_i x_1 \ldots x_m}$ – коэффициент детерминации для зависимости фактора x_i со всеми другими факторами уравнения множественной регрессии;

$n - m - 1$ – число степеней свободы для остаточной суммы квадратов отклонений.

Как видим, чтобы воспользоваться данной формулой, необходимы матрица межфакторной корреляции и расчет по ней соответствующих коэффициентов детерминации $R^2_{x_i x_1 \ldots x_m}$. Так, для уравнения $y = a + b_1 x_1 + b_2 x_2 + b_3 x_3$ оценка значимости коэффициентов регрессии b_1, b_2, b_3 предполагает расчет трех межфакторных коэффициентов детерминации: $R^2_{x_1 \cdot x_2 x_3}$, $R^2_{x_2 \cdot x_1 x_3}$, $R^2_{x_3 \cdot x_1 x_2}$.

Взаимосвязь показателей частного коэффициента корреляции, частного F-критерия и t-критерия Стьюдента для коэффициентов чистой регрессии может использоваться в процедуре отбора факторов. Отсев факторов при построении уравнения регрессии методом исключения практически можно осуществлять не только по частным коэффициентам корреляции, исключая на каждом шаге фактор с наименьшим незначимым значением частного коэффициента корреляции, но и по величинам t_{b_i} и F_{x_i}. Частный F-критерий широко используется и при построении модели методом включения переменных и шаговым регрессионным методом.

3. **Контрольные вопросы**

1. В чем состоит спецификация модели множественной регрессии?
2. Отбор факторов при построении уравнения множественной регрессии.
3. Сформулируйте требования, предъявляемые к факторам, для включения их в модель множественной регрессии.
4. Оценка параметров уравнения множественной регрессии.
5. Как интерпретируются коэффициенты регрессии линейной модели?
6. Множественная корреляция.
7. К каким трудностям приводит мультиколлинеарность факторов, включенных в модель, и как они могут быть преодолены?
8. Какие коэффициенты используются для оценки сравнительной силы воздействия факторов на результат?
9. Частные коэффициенты корреляции.
10. F-критерий Фишера и частный F-критерий Фишера для уравнения множественной регрессии.
11. t-критерий Стьюдента для уравнения множественной регрессии.

4. Построение модели множественной регрессии.

Рассмотрим **пример 1** (для сокращения объема вычислений ограничимся только десятью наблюдениями). Акцент сделаем на спецификации модели, сосредоточив внимание на отборе факторов в модель.

Пусть имеются следующие данные (условные) о сменной добыче угля на одного рабочего y (т), мощности пласта x_1 (м) и уровне механизации работ x_2 (%), характеризующие процесс добычи угля в 10 шахтах. Данные 10 наблюдений представлены в таблице 2.2.

Таблица 2.2

№	1	2	3	4	5	6	7	8	9	10
x_1	8	11	12	9	8	8	9	9	8	12
x_2	5	8	8	5	7	8	6	4	5	7
y	5	10	10	7	5	6	6	5	6	8

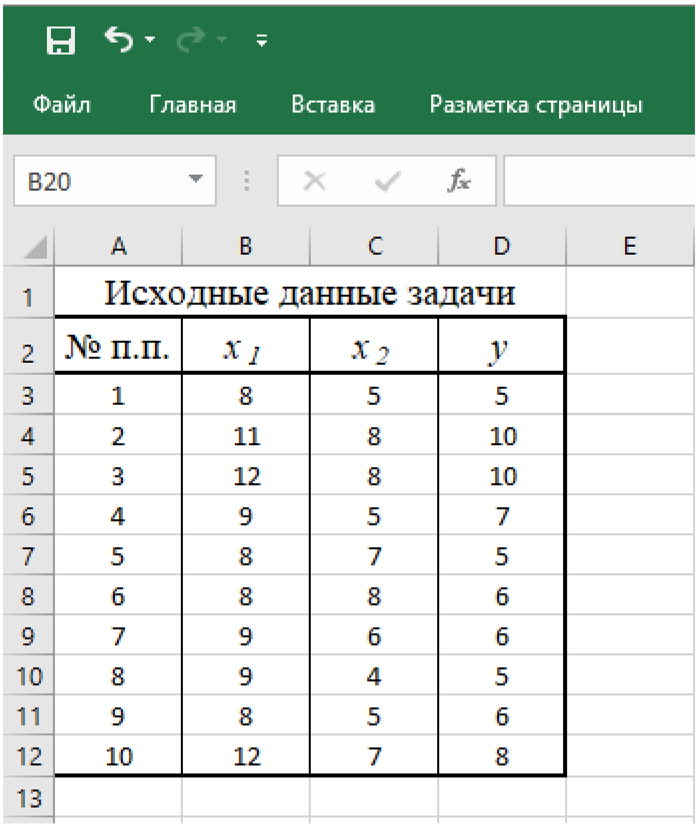
Предполагая, что между переменными y, x_1, x_2 существует линейная корреляционная зависимость, найдем уравнение регрессии y по x_1 и x_2:

$$y = a + b_1 \cdot x_1 + b_2 \cdot x_2 + \varepsilon,$$

где ($\varepsilon = y - y_x$).

Определить параметры линейного уравнения данного вида можно с использованием функции «Регрессия» надстройки «Анализ данных» табличного процессора Excel.

Прежде всего, надо сформировать таблицу с исходными данными для анализа на листе Excel. Скриншот листа Excel с расположенной на нем таблицей исходных данных представлен на рис. 1.

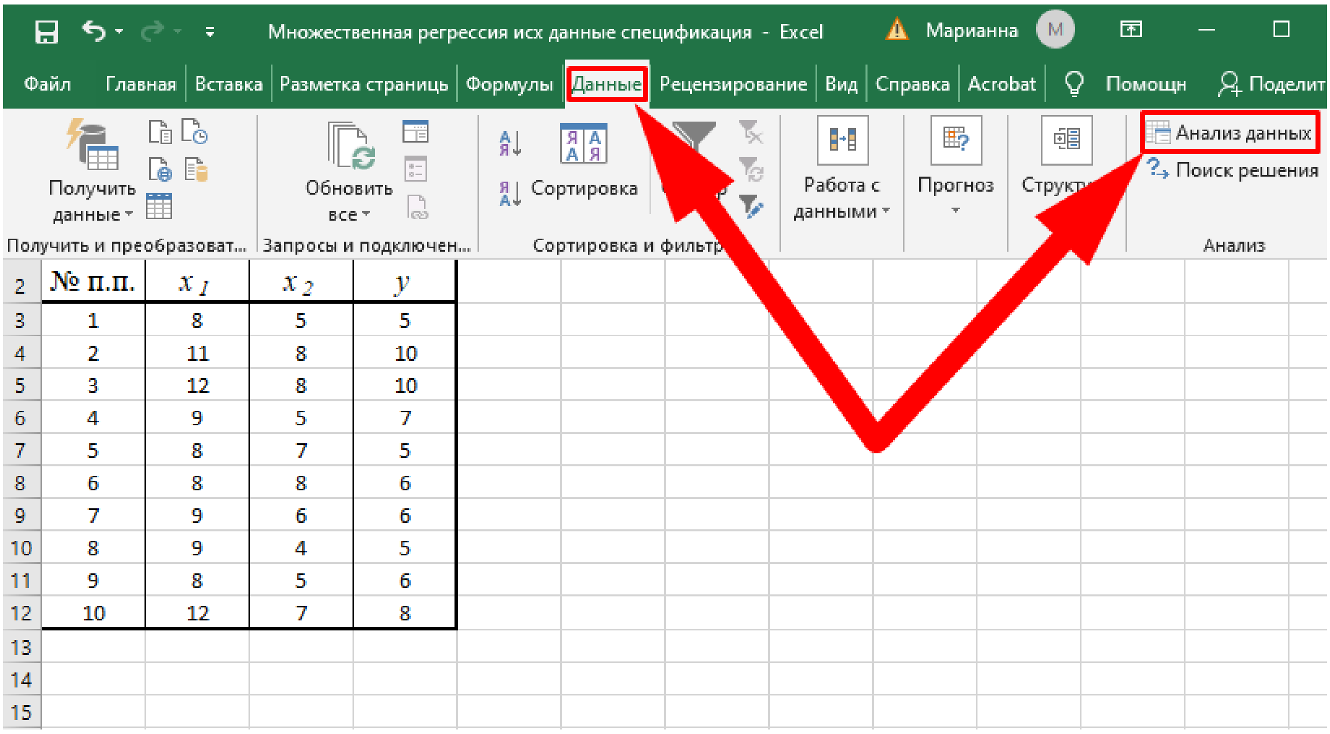


Исходные данные задачи			
№ п.п.	x_1	x_2	y
1	8	5	5
2	11	8	10
3	12	8	10
4	9	5	7
5	8	7	5
6	8	8	6
7	9	6	6
8	9	4	5
9	8	5	6
10	12	7	8

Рис. 1. Скриншот листа Excel с расположенной на нем таблицей исходных данных для решения задачи.

Следует обратить внимание на расположение числовых данных: они должны быть введены в виде столбцов, а не строк, как было представлено условие задачи в табл. 2.2.

Для оценки параметров уравнения линейной регрессии $\hat{y} = a + b_1 \cdot x_1 + b_2 \cdot x_2$ можно воспользоваться методом наименьших квадратов, который реализуется функцией «Регрессия» надстройки «Анализ данных» табличного процессора Excel. На рисунке 2 стрелками указаны пункты меню, к которым нужно обратиться, чтобы воспользоваться функцией **Регрессия** табличного процессора Excel.



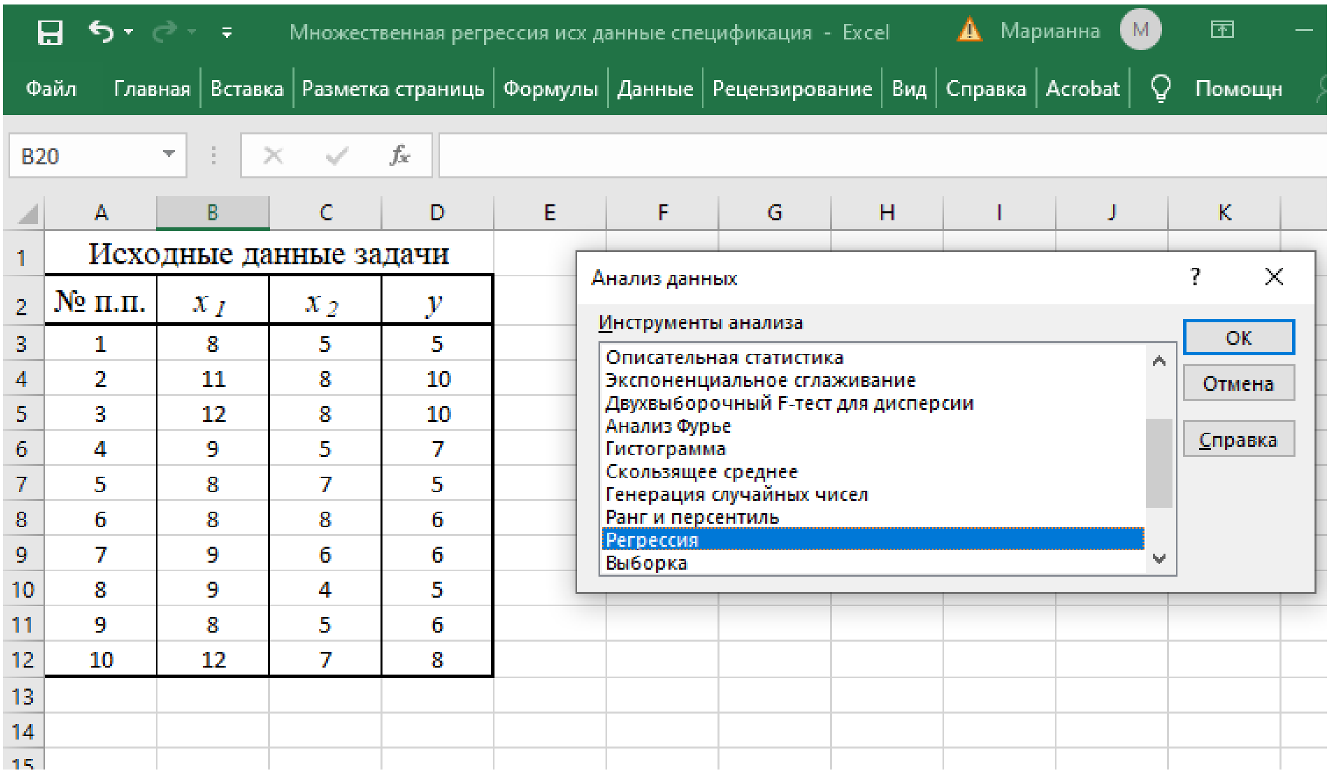
№ п.п.	x_1	x_2	y
1	8	5	5
2	11	8	10
3	12	8	10
4	9	5	7
5	8	7	5
6	8	8	6
7	9	6	6
8	9	4	5
9	8	5	6
10	12	7	8

Рис. 2. Скриншот листа Excel с выделенными пунктами меню.

Следует заметить, что пакет анализа должен быть установлен на компьютере. Если же это не было сделано ранее, нужно его установить. Для этого можно зайти во вкладку **Надстройки (Файл – Параметры – Надстройки – Надстройки Excel - Перейти**), после чего требуется установить флажок для **Пакета анализа**.

После вызова **Анализа данных** на экране появляется окно с перечнем инструментов анализа. Передвигая бегунок, следует найти требуемую функцию, выделить ее и выбрать, нажав кнопку **ОК**.

Скриншот листа Excel с окном **Анализ данных** и выбранной функцией **Регрессия** представлен на рисунке 3.



Исходные данные задачи

№ п.п.	x_1	x_2	y
1	8	5	5
2	11	8	10
3	12	8	10
4	9	5	7
5	8	7	5
6	8	8	6
7	9	6	6
8	9	4	5
9	8	5	6
10	12	7	8

Рис. 3. Скриншот листа Excel с окном «Анализ данных».

После вызова режима **Регрессия** на экране появляется диалоговое окно. Вид диалогового окна **Регрессия** представлен на рис. 4.

В диалоговом окне задаются следующие параметры:

1. *Входной интервал Y* – вводится диапазон адресов ячеек, содержащих значения y_i, т.е. зависимой переменной или признака-результата (ячейки должны составлять один столбец).
2. *Входной интервал X* – вводится диапазон адресов ячеек, содержащих значения независимых переменных или признаков-факторов. Значения каждой переменной представляются одним столбцом. В режиме **Регрессия** можно построить не только линейную парную, но и множественную регрессию. Количество переменных может быть не более 16.

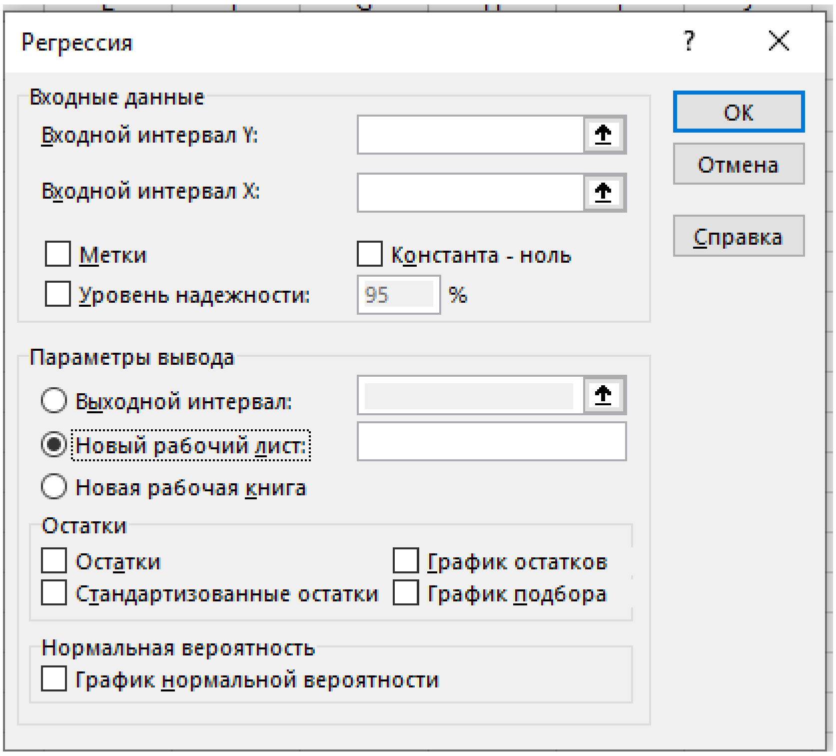


Рис. 4. Скриншот диалогового окна **Регрессия.**

3. *Метки* – включается, если первая строка во входном диапазоне содержит заголовок, чтобы программа не трактовала его как числовые данные. В этом случае автоматически будут созданы стандартные названия.
4. *Уровень надежности* – при включении этого параметра задается надежность при построении доверительных интервалов.
5. *Константа-ноль* – при включении этого параметра коэффициент *a* (свободный член регрессии) равен 0.
6. *Выходной интервал* – при включении активизируется поле, куда необходимо ввести адрес левой верхней ячейки выходного диапазона, который будет содержать ячейки с результатами вычислений режима **Регрессия**.
7. *Новый рабочий лист* - при включении этого параметра открывается новый лист, в который, начиная с ячейки A1, вставляются результаты работы режима **Регрессия.**

8. *Новая рабочая книга* – при включении этого параметра открывается новая книга, на первом листе которой, начиная с ячейки A1, вставляются результаты работы режима **Регрессия**.
9. *Остатки* – при включении вычисляется столбец, содержащий остатки для всех точек наблюдений (исходных данных) $y_i - \hat{y}_i, i = 1,...,n$.
10. *Стандартизованные остатки* – при включении вычисляется столбец, содержащий стандартизованные остатки.
11. *График остатков* – при включении выводятся точечные графики остатков $y_i - \hat{y}_i, i = 1,...,n$, в зависимости от значений переменных x_j, $j=1,...,m$. Количество графиков равно числу m переменных x_j.
12. *График подбора* – при включении выводятся точечные графики предсказанных по построенной регрессии значений $\hat{y}_i$ от значений переменных x_j, при $j=1,...,m$. Количество графиков равно числу m переменных x_j.

    Для парной регрессии $m=1$, поскольку признак-фактор в модели (уравнении регрессии) только один. Будет построен один график подбора и один график остатков.

В диалоговом окне **Регрессия** нужно сделать следующее:

- ввести в окне редактирования **Входной интервал Y** диапазон зависимой переменной, что можно сделать, выделив нужный столбец y_i таблицы исходных данных на листе Excel;

Скриншот диалогового окна **Регрессия** после ввода диапазона ячеек со значениями зависимой переменной y путем выделения соответствующего столбца таблицы исходных данных представлен на рисунке 5.

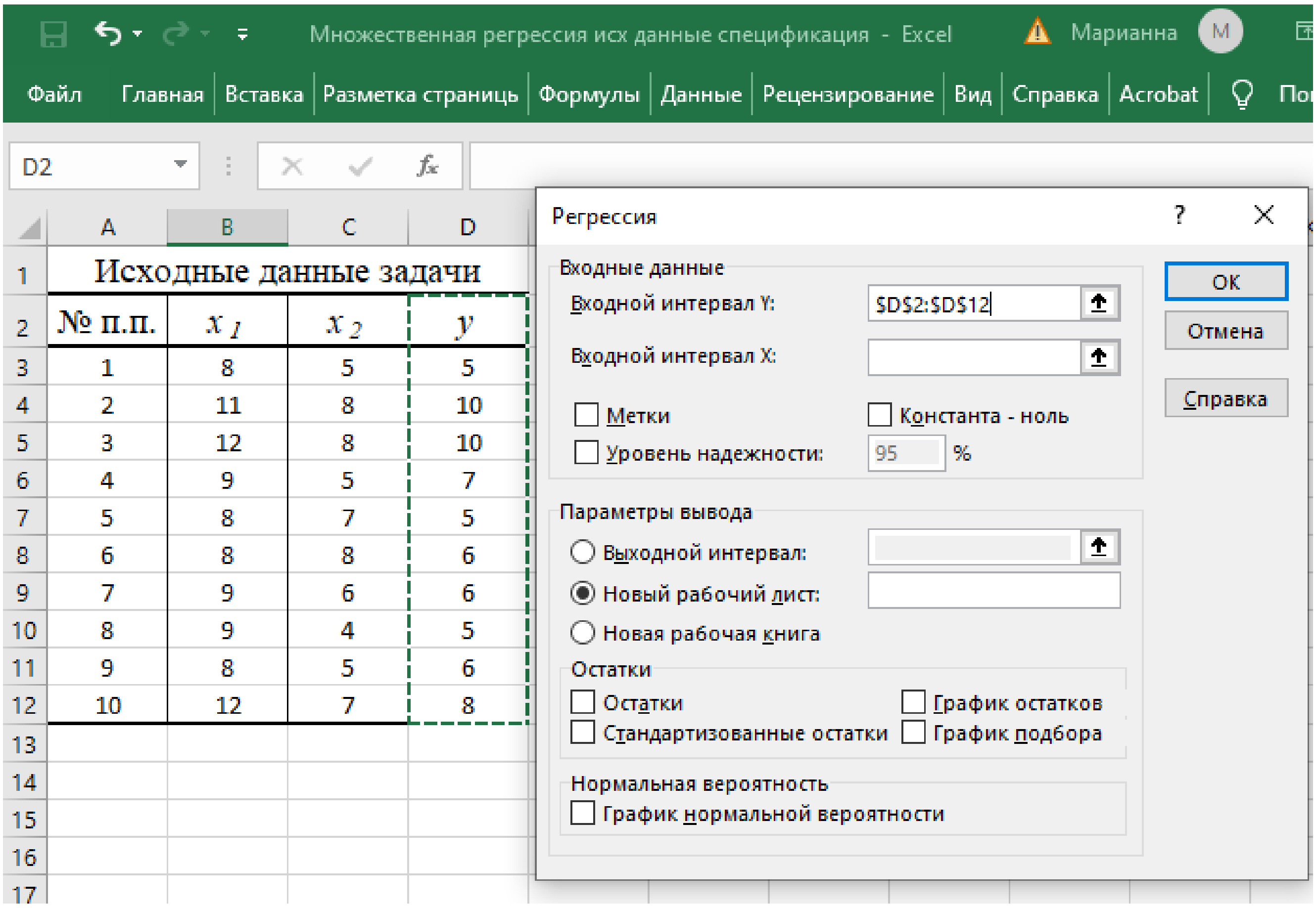


Рис. 5. Скриншот диалогового окна **Регрессия** после ввода диапазона ячеек со значениями зависимой переменной *y*.

- ввести в окне редактирования **Входной интервал X** диапазон факторных переменных x_i (т.е. x_1 и x_2, например, выделив весь диапазон ячеек от верхнего левого угла до нижнего правого угла прямоугольника, включающего оба столбца x_1 и x_2, т.е. от ячейки B3 до ячейки C12).

На рисунке 6 представлен скриншот вида листа Excel с диалоговым окном **Регрессия** после ввода диапазона ячеек со значениями независимых переменных x_1 и x_2.

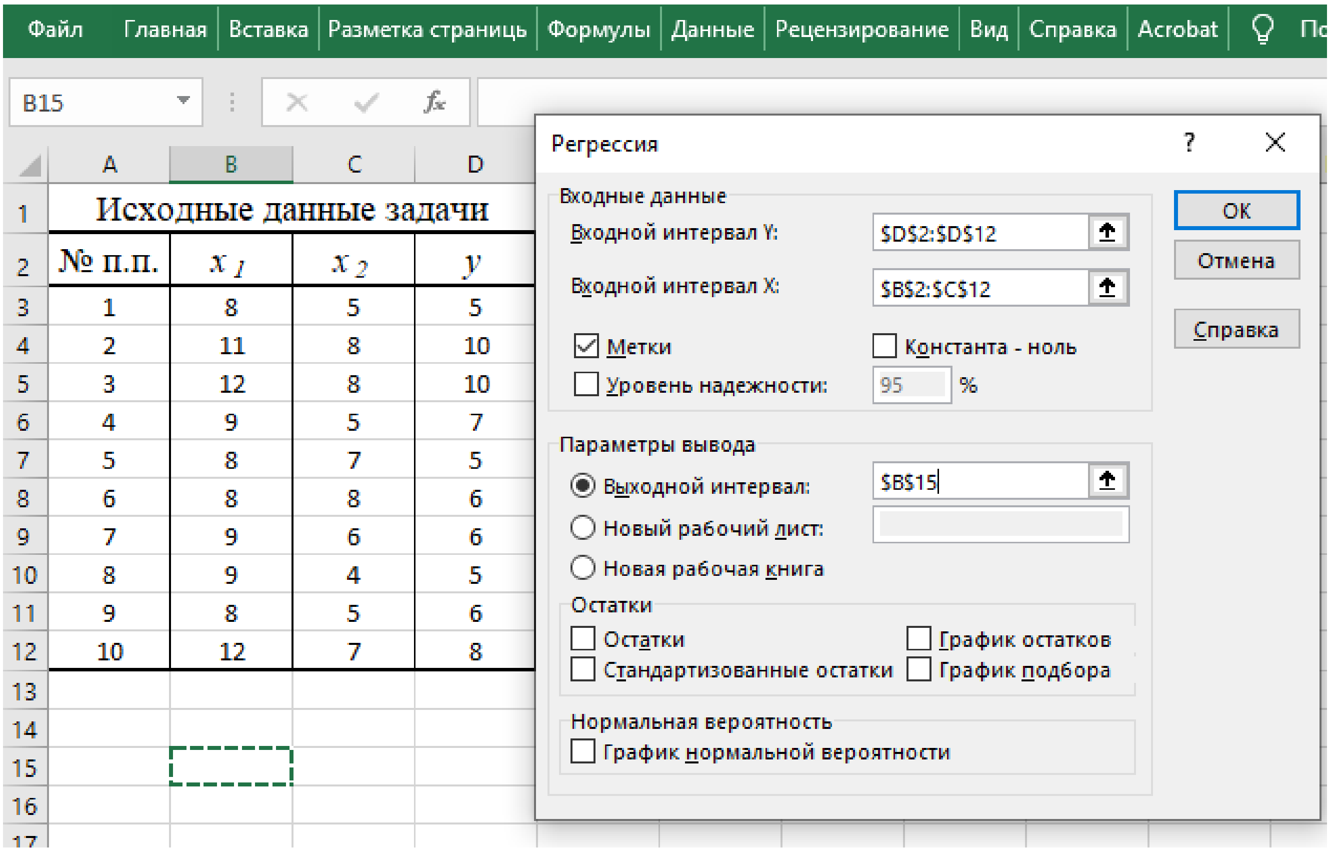


Рис. 6. Скриншот диалогового окна **Регрессия** после ввода диапазона ячеек со значениями независимых переменных.

-установить флажок **Метки**, если первая строка содержит название столбцов;

Если в диалоговом окне введены только указания на ячейки, содержащие данные, флажок **Метки** не должен быть установлен. На рис.6, иллюстрирующем решение данной задачи, флажок **Метки** установлен, поскольку первая строка содержит название столбцов.

-установить флажок **Константа-ноль**, если в уравнении регрессии отсутствует свободный член *a*; В данной задаче не предполагается равенство нулю свободного члена регрессии, поэтому флажок **Константа-ноль** не должен быть установлен.

- ввести в окне редактирования **Выходной интервал** номер свободной ячейки на рабочем столе (или выбрать новый лист);

На рис. 6 приведен скриншот диалогового окна **Регрессия**, где в параметрах вывода установлен флажок у поля **Выходной интервал**, следовательно вывод результатов будет выполнен на данный лист, т.е. Excel поместит таблицы итогов работы, начиная с заданной ячейки на текущем листе (левой верхней всего диапазона).

- нажать **ОК**.

Таблицы итогов работы режима **Регрессия** приведены для данной задачи на рис. 7.

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественны	0,90089922
R-квадрат	0,811619404
Нормированны	0,757796377
Стандартная ош	0,950908439
Наблюдения	10

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	2	27,27041	13,63521	15,07941	0,002902
Остаток	7	6,329588	0,904227		
Итого	9	33,6			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	-3,539325843	1,906581	-1,85637	0,105773	-8,04767	0,969021
x1	0,853932584	0,220504	3,872634	0,006111	0,332523	1,375342
x2	0,367041199	0,242948	1,510779	0,174596	-0,20744	0,941523

Рис. 7. Скриншот результатов работы режима **Регрессия**.

Заметим, что таблицы, которые являются результатом работы режима **Регрессия**, являются такими же таблицами, как и все другие таблицы Excel. Это значит, что работать с ними можно так же, как и с другими, меняя шрифт, цвет, размер и многое другое, например, приведя их в более удобный и привычный для восприятия вид в соответствии с потребностями или желаниями.

Дадим краткую интерпретацию показателям, значения которых вычисляются в режиме **Регрессия**.

а) *Регрессионная статистика*:

Сначала рассмотрим показатели, объединенные названием «Регрессионная статистика». Для удобства эта таблица представлена отдельно на рис. 8.

Регрессионная статистика	
Множественный R	0,90089922
R-квадрат	0,811619404
Нормированный R-квадрат	0,757796377
Стандартная ошибка	0,950908439
Наблюдения	10

Рис. 8. Таблица «Регрессионная статистика»

Значения показателей, необходимых для решения задачи, округлим до трех знаков после запятой.

- Множественный R – для данной задачи – это индекс (коэффициент) множественной корреляции $R_{yx_1x_2}$ = 0,901. Также он получил название линейного коэффициента множественной корреляции, или, что то же самое, совокупного коэффициента корреляции;

- R-квадрат- индекс (коэффициент) детерминации $R^2_{yx_1x_2}$ = 0,812;

- Нормированный R-квадрат - приведенный коэффициент детерминации R^2 = 0,758

- Стандартная ошибка – оценка Se для среднеквадратического отклонения б .

- наблюдения – число наблюдений n = 10;

Значение линейного коэффициента множественной корреляции говорит о высокой тесноте связи между результативным признаком и факторными.

Значение коэффициента детерминации можно выразить в процентах. Он показывает, что на 81% вариация результативного признака *y* определяется (объясняется) вариацией факторных признаков (x_1, x_2) в этой модели.

б) *Дисперсионный анализ*:

Показатели, объединенные названием «Дисперсионный анализ» для удобства представлены на рис. 9.

Дисперсионный анализ					
	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	2	27,2704	13,6352	15,0794	0,0029
Остаток	7	6,32959	0,90423		
Итого	9	33,6			

Рис. 9. Таблица «Дисперсионный анализ»

- столбец *df* – число степеней свободы для уравнения регрессии (строка **Регрессия**), для остаточной вариации (строка **Остаток**), и общая вариация (строка **Итого**). Для строки *Регрессия* показатель равен *m* – числу параметров при независимых переменных (x_1, x_2 в данной задаче, т.е. 2); для строки *Остаток* – равен *n-m-1* (т.е. 7); для строки *Итого* (общая дисперсия) – *n-1* или числу наблюдений без единицы, т.е. 9;

- столбец *SS* содержит суммы квадратов отклонений: сумму квадратов отклонений теоретических данных от среднего значения (строка **Регрессия**), сумму квадратов отклонений фактических данных от теоретических (строка **Остаток**) и сумму квадратов отклонений фактических данных от среднего значения (строка

Итого);

- в столбце *MS* показаны дисперсии на одну степень свободы: объясненная (факторная) дисперсия (для строки **Регрессия**) и остаточная дисперсия (для строки **Остаток**);

- в столбце ***F*** показано расчетное значение *F*-критерия Фишера, которое сравнивают с табличным $F_{\text{табл}}(\alpha;\, k_1;\, k_2)$ при уровне значимости α и степенях свободы $k_1 = m$ и $k_2 = n - m - 1$. При этом, если фактическое значение F-критерия больше табличного, то признается статистическая значимость уравнения в целом. Таблица значений *F*-критерия Фишера при уровне значимости $\alpha = 0{,}05$ приведена в Приложении 1;

Для нашей задачи, т.е. при α=0,05, k_1=2, k_2=10-1-1=7 $F_{табл}$ равно 4,74. Расчетное значение F-критерия равно 15,079, что значительно превышает табличное, т.е. следует признать статистическую значимость уравнения регрессии в целом.

- в столбце **значимость *F*** показано значение уровня значимости, соответствующее вычисленной величине *F*-критерия и равное вероятности того, что расчетное значение *F*-критерия меньше или равно табличному. Если вероятность меньше уровня значимости α (обычно $\alpha = 0{,}05$), то построенная регрессия является значимой;

Значимость *F* рассчитанная на порядок меньше уровня значимости α = 0,05, следовательно построенная регрессия является значимой.

в) *Перейдем к следующей группе показателей*, объединенных в таблице, показанной на рис. 10.

	Коэффи-циенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	-3,5393	1,90658	-1,85637	0,10577	-8,04767	0,96902	-8,04767	0,96902
x1	0,85393	0,2205	3,87263	0,00611	0,33252	1,37534	0,33252	1,37534
x2	0,36704	0,24295	1,51078	0,1746	-0,20744	0,94152	-0,20744	0,94152

Рис. 10. Продолжение таблицы с результатами работы режима ***Регрессия***

Приведенная на рис. 10 таблица включает, кроме оценок параметров, также их среднеквадратические ошибки, вероятности ошибочного решения (P-значение), нижние и верхние интервальные оценки параметров с вероятностью 95%.

В столбце *Коэффициенты* показаны значения коэффициентов уравнения регрессии. В строке *Y-пересечение* – представлено значение параметра a, в строке *Переменная x1* – значение параметра b_1, в строке *Переменная x2* – значение параметра b_2.

В данном случае это свободный член и коэффициенты регрессии для следующего уравнения:

$$\hat{y} = a + b_1 x_1 + b_2 x_2$$

Продолжим рассмотрение показателей из группы, представленной в таблице на рис. 10.

В столбце *Стандартная ошибка* приведены значения стандартных ошибок для параметров регрессии;

В столбце *t-статистика* – значения статистик Стьюдента, рассчитанные для соответствующих параметров регрессии. Фактические (расчетные) значения сравнивают с табличными, если фактическое значение больше табличного, то признается статистическая значимость параметра;

В столбце *Р-значение* содержатся вероятности случайных событий непревышения расчетной статистикой Стьюдента для соответствующего параметра регрессии табличного значения. Если эта вероятность меньше уровня значимости α (обычно задается $\alpha = 0,05$), то принимается гипотеза о значимости соответствующего коэффициента регрессии.

Табличное значение *t*-критерия Стьюдента находят по таблице, при $\alpha = 0,05$ и числе степеней свободы $n-2$. Таблица критических значений t -критерия Стьюдента при уровне значимости 0,10; 0,05; 0,01 приведена в Приложении 1;

Для рассматриваемой задачи $n-2 = 8$, $t_{крит}$ = 2,306, т.е. расчетные значения *t* – статистики для двух параметров меньше критического (табличного) значения, следовательно, отклоняется гипотеза о значимости обоих параметров регрессии.

Столбцы *Нижние 95%* и *Верхние 95%* - содержат соответственно нижние и верхние границы интервалов для оцениваемых параметров a, b_1 и b_2.

Следует заметить, что число наблюдений для модели с двумя факторами должно быть в полтора - два раза больше, чем в рассматриваемом варианте. Этим в том числе объясняется статистическая незначимость параметров полученной регрессии. Однако, рассмотреть приемы и способы отбора факторов в модель, а также оценки их значимости, можно на примере с небольшим количеством наблюдений.

Подставим найденные МНК значения для параметров в уравнение регрессии:

$$\hat{y} = -3,539 + 0,854 \cdot x_1 + 0,367 \cdot x_2$$

Данное уравнение показывает, что при увеличении только мощности пласта x_1 (при неизменном x_2) на 1 м добыча угля на одного рабочего y увеличится в среднем на 0,854 т, а при увеличении только уровня механизации работ x_2 (при неизменном x_1) на 1% – в среднем на 0,367 т.

Перейдем к уравнению регрессии в стандартизованном масштабе:

$$t_y = \beta_1 t_{x_1} + \beta_2 t_{x_2} + \varepsilon,$$

Стандартизованные коэффициенты регрессии показывают, на сколько единиц изменится в среднем результат, если соответствующий фактор x_i изменится на одну единицу при неизменном среднем уровне других факторов. В силу того, что все переменные заданы как центрированные и нормированные, стандартизованные коэффициенты регрессии β_i можно сравнивать между собой. Сравнивая их друг с другом, можно ранжировать факторы по силе их воздействия на результат. В этом основное достоинство стандартизованных коэффициентов регрессии в отличие от коэффициентов «чистой» регрессии, которые несравнимы между собой.

Воспользуемся формулой 2.6, которая выражает связь коэффициентов «чистой» регрессии и стандартизованных коэффициентов регрессии:

$$b_i = \beta_i \frac{\sigma_y}{\sigma_{x_i}},$$

следовательно

$$\beta_i = b_i \frac{\sigma_{x_i}}{\sigma_y}$$

Для расчета стандартизованных коэффициентов регрессии необходимы значения среднеквадратического отклонения для всех переменных. Можно воспользоваться статистической функцией Excel КВАДРОТКЛ(), которая возвращает сумму отклонений точек данных от среднего по выборке.

Скриншот окна **Вставка функции** Excel приведен на рис. 11.

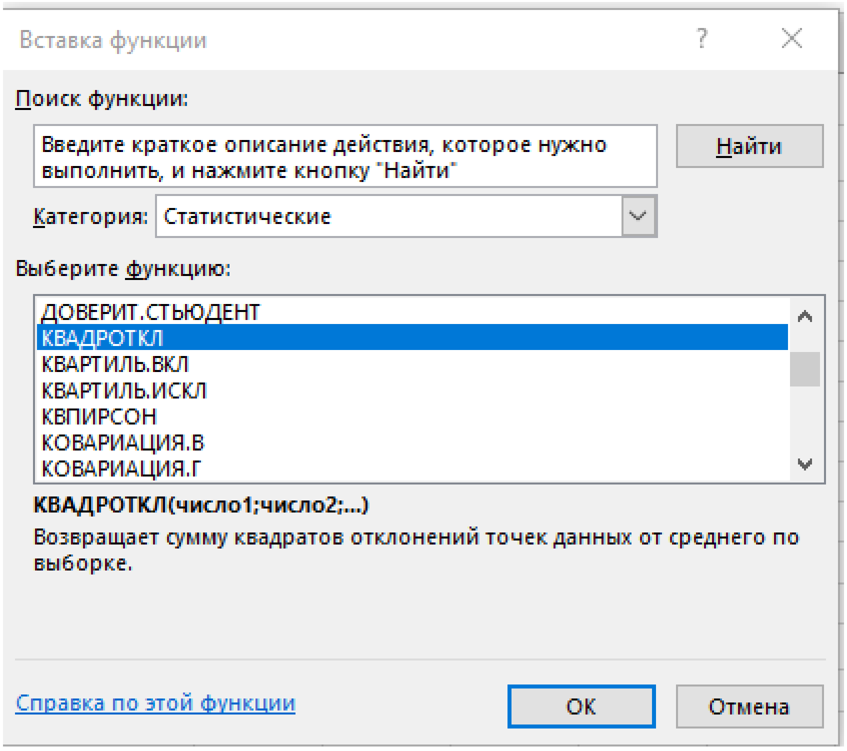


Рис. 11. Скриншот окна **Вставка функции** при обращении к функции КВАДРОТКЛ().

Функция имеет следующий синтаксис: КВАДРОТКЛ(число1;[число2];...).

Аргумент функции КВАДРОТКЛ ***число1*** является обязательным, последующие числа - нет. Таким образом, задавая в качестве массива данных столбцы значений переменных, мы получим на выходе суммы квадратов отклонений от среднего значения соответствующих переменных, разделив которые на число наблюдений, получим дисперсии для этих переменных.

Скриншот окна функции КВАДРОТКЛ представлен на рис.12.

Скриншот листа Excel с формулой вычисления дисперсии независимой переменной x_1 в ячейке **B14** представлен на рис.13.

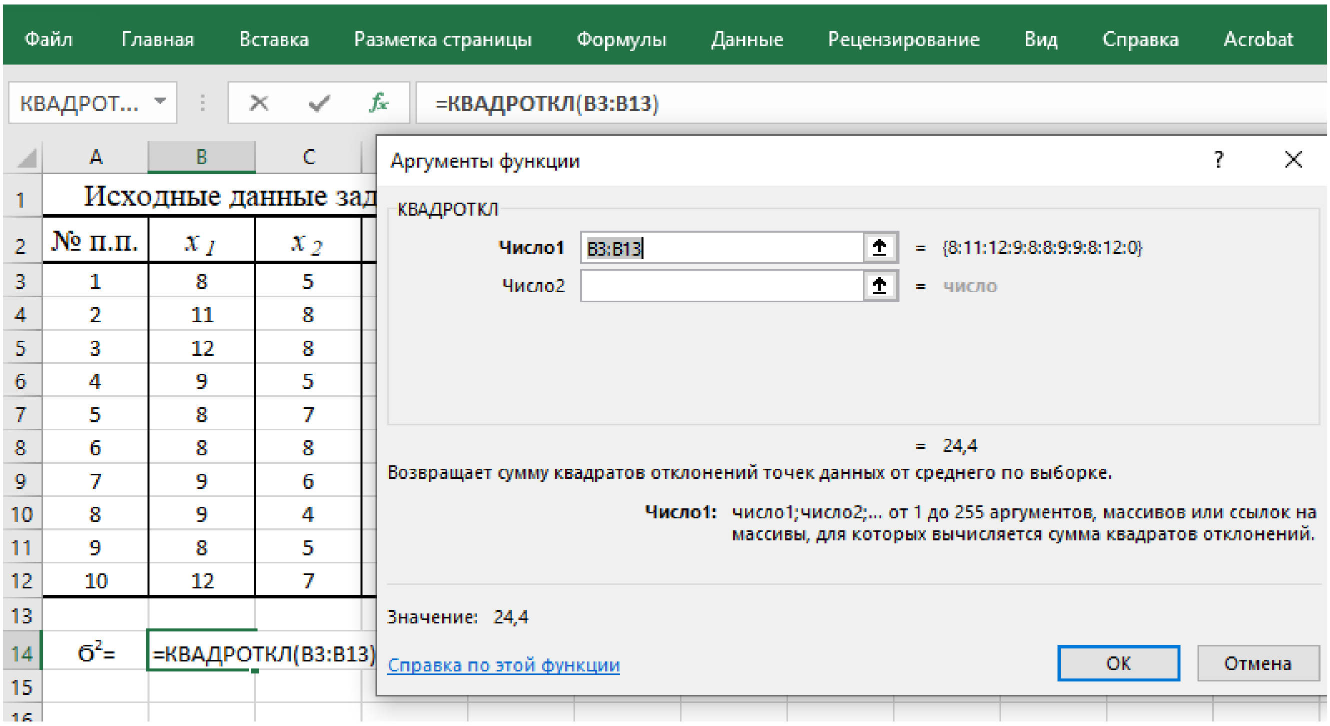
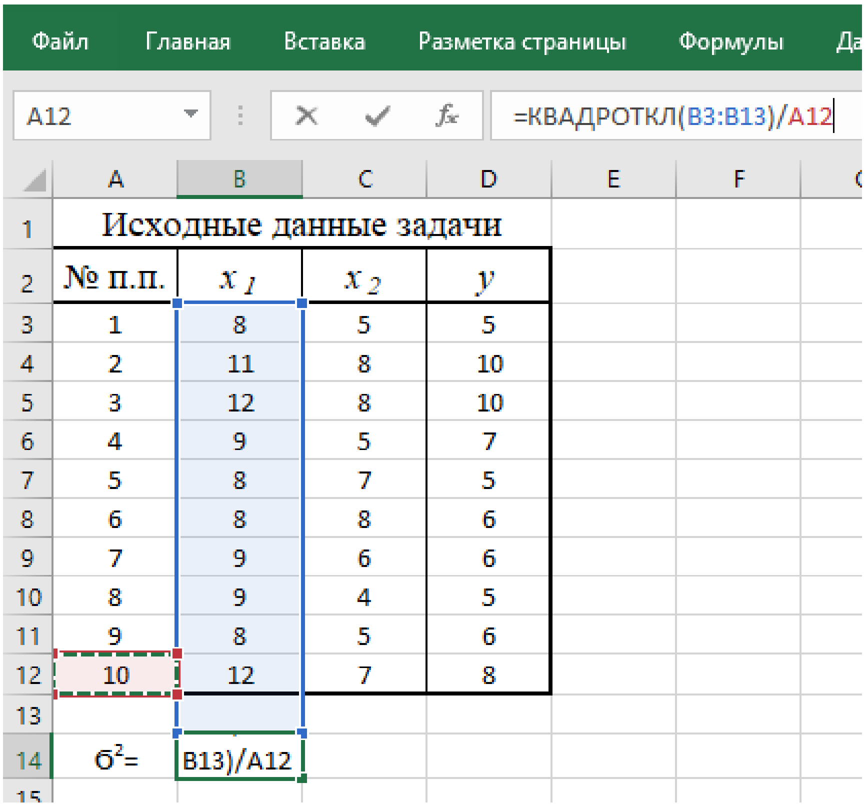


Рис. 12. Скриншот окна функции КВАДРОТКЛ().



№ п.п.	x_1	x_2	y
1	8	5	5
2	11	8	10
3	12	8	10
4	9	5	7
5	8	7	5
6	8	8	6
7	9	6	6
8	9	4	5
9	8	5	6
10	12	7	8

$б^2$= B13)/A12

Рис. 13. Скриншот листа Excel с формулой для вычисления дисперсии x_1.

Среднеквадратическое отклонение легко получить в EXCEL, извлекая квадратный корень из дисперсии с помощью математической функции КОРЕНЬ.

Скриншот окна функции КОРЕНЬ представлен на рис. 14.

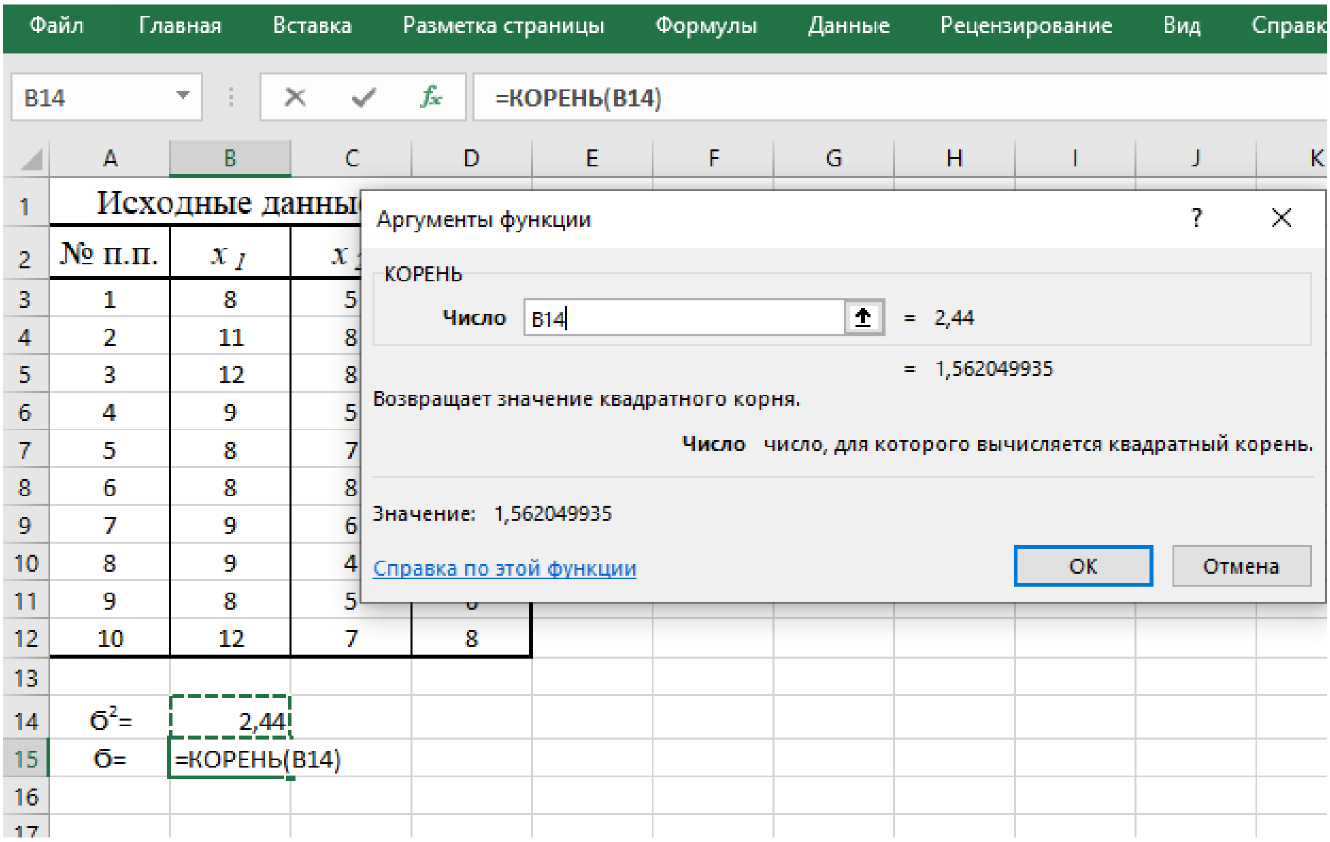


Рис. 14. Скриншот окна функции КОРЕНЬ.

Результаты расчетов стандартизованных коэффициентов регрессии в Excel представлены в таблице на рис.15.

Стандартизованные коэффициенты регрессии:

$$\beta_1 = b_1 \frac{\sigma_{x_1}}{\sigma_y} = 0,8539 \cdot \frac{1,56205}{1,83303} = 0,72769,$$

$$\beta_2 = b_2 \frac{\sigma_{x_2}}{\sigma_y} = 0,367 \cdot \frac{1,4177}{1,83303} = 0,28389.$$

	A	B	C	D	E
1	Исходные данные задачи				
2	№ п.п.	x_1	x_2	y	
3	1	8	5	5	
4	2	11	8	10	
5	3	12	8	10	
6	4	9	5	7	
7	5	8	7	5	
8	6	8	8	6	
9	7	9	6	6	
10	8	9	4	5	
11	9	8	5	6	
12	10	12	7	8	
13					
14	σ^2	2,44	2,01	3,36	
15	σ	1,56205	1,417745	1,83303	
16	β_1=	0,727694			
17	β_2=	0,283885			
18					

Рис. 15. Скриншот результатов расчетов стандартизованных коэффициентов регрессии.

Т.е. уравнение будет выглядеть следующим образом:

$$\hat{t}_y = 0,728 \cdot t_{x_1} + 0,284 \cdot t_{x_2} .$$

Так как стандартизованные коэффициенты регрессии можно сравнивать между собой, то можно сказать, что мощность пласта оказывает большее влияние на сменную добычу угля, чем уровень механизации работ.

Сравнивать влияние факторов на результат можно также при помощи средних коэффициентов эластичности (2.11):

$$\bar{Э}_i = b_i \cdot \frac{\bar{x}_i}{\bar{y}_{x_i}} .$$

Для расчета средних значений переменных удобно воспользоваться функцией Excel СРЗНАЧ(), которая возвращает среднее арифметическое своих аргументов, которые могут быть числами, массивами и т.д.

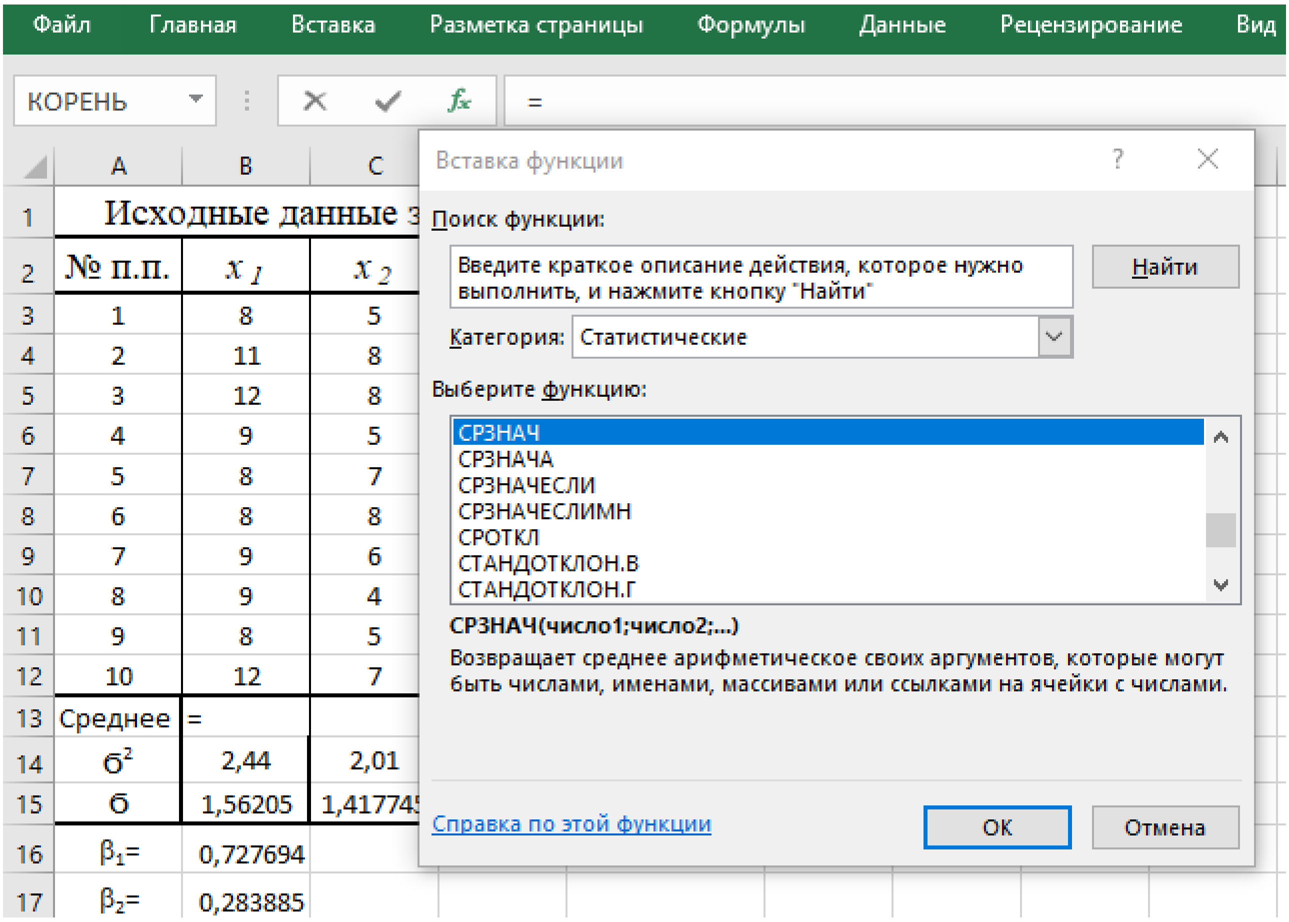


Рис. 16. Скриншот окна **Вставка функции** при обращении к функции СРЗНАЧ.

Обязательным аргументом функции **СРЗНАЧ**(число1;[число2];…) является ***число1***. Он может быть первым числом, ссылкой на ячейку или диапазон, для которого требуется вычислить среднее значение. Для нашей задачи надо выделить диапазон значений переменной, для которой требуется вычислить среднее значение, т.е. столбец значений.

Аргумент ***число2*** - необязательный.

Скриншот окна функции **СРЗНАЧ** для расчета среднего значения переменной x_1 приведен на рис.17.

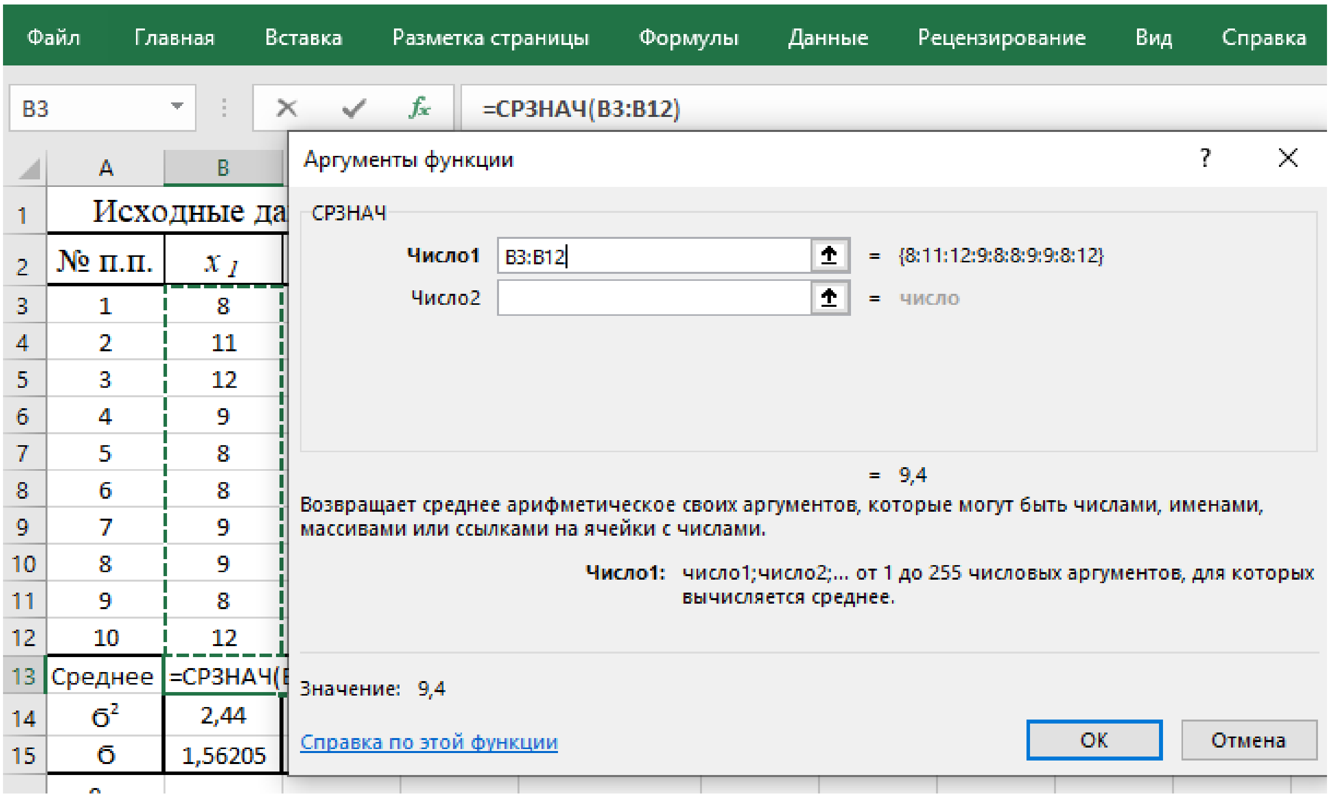


Рис. 17. Скриншот окна функции СРЗНАЧ при расчете среднего значения переменной x_1.

Скриншот листа Excel с таблицей, где рассчитаны средние значения для зависимой и независимых переменных (y, x_1, x_2), необходимых для вычисления средних коэффициентов эластичности, а также значений самих коэффициентов приведен на рис. 18.

Вычисляем:

$$\bar{Э}_1 = 0,8539 \cdot \frac{9,4}{6,8} = 1,1804 \qquad \bar{Э}_2 = 0,367 \cdot \frac{6,3}{6,8} = 0,3401, \quad .$$

1	Исходные данные задачи			
2	№ п.п.	x_1	x_2	y
3	1	8	5	5
4	2	11	8	10
5	3	12	8	10
6	4	9	5	7
7	5	8	7	5
8	6	8	8	6
9	7	9	6	6
10	8	9	4	5
11	9	8	5	6
12	10	12	7	8
13	Среднее	9,4	6,3	6,8
14	б^2	2,44	2,01	3,36
15	б	1,56205	1,417745	1,83303
16	β_1=	0,727694		
17	β_2=	0,283885		
18	$\bar{Э}_1 =$	1,180436		
19	$\bar{Э}_2 =$	0,340053		

Рис. 18. Скриншот окна листа Excel со значениями средних коэффициентов эластичности.

Таким образом, увеличение только мощности пласта (от своего среднего значения) или только уровня механизации работ на 1% увеличивает в среднем сменную добычу угля на 1,18% или 0,34% соответственно. Таким образом, подтверждается большее влияние на результат y фактора x_1, чем фактора x_2.

Построим далее матрицу парных коэффициентов корреляции для зависимости $y = \hat{f}(x_1, x_2)$.

Матрица коэффициентов корреляции в Excel строится с помощью инструмента «**Корреляция**» из пакета «**Анализ данных**». На вкладке «**Данные**» в группе «**Анализ**» открываем пакет «**Анализ данных**». В списке инструментов анализа выбираем «**Корреляция**». Нажимаем **ОК**.

Скриншот окна **Анализ данных** с выбранным инструментом анализа «**Корреляция**» приведен на рис. 19.

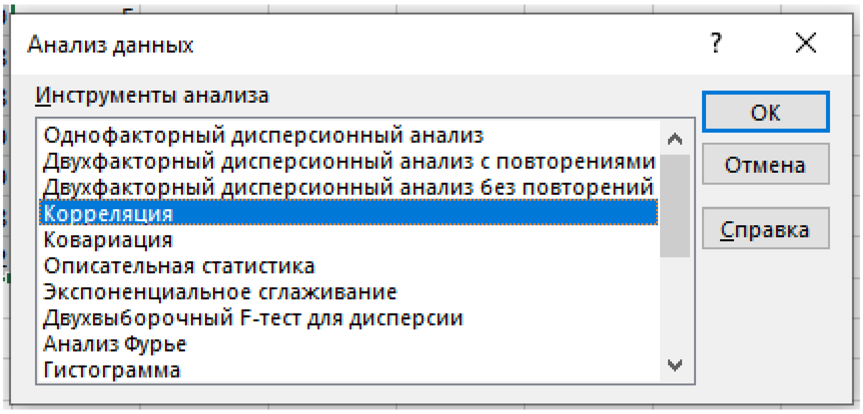


Рис. 19. Скриншот окна **Анализ данных** с выбранным инструментом анализа «**Корреляция**».

Вид окна функции Корреляция приведен на рис. 20.

Задаем параметры для анализа данных.

Входной интервал – диапазон ячеек со значениями. Для того, чтобы корреляционная матрица имела привычный и удобный вид, порядок столбцов удобно использовать именно приведенный в таблице со скриншота (слева – направо сохранить порядок: y, x_1, x_2). Тогда корреляционная матрица будет иметь такой вид как на рис. 21.

Группирование – по столбцам (анализируемые данные сгруппированы в столбцы).

Выходной интервал – ссылка на ячейку, с которой начнется построение матрицы. Размер диапазона определится автоматически.

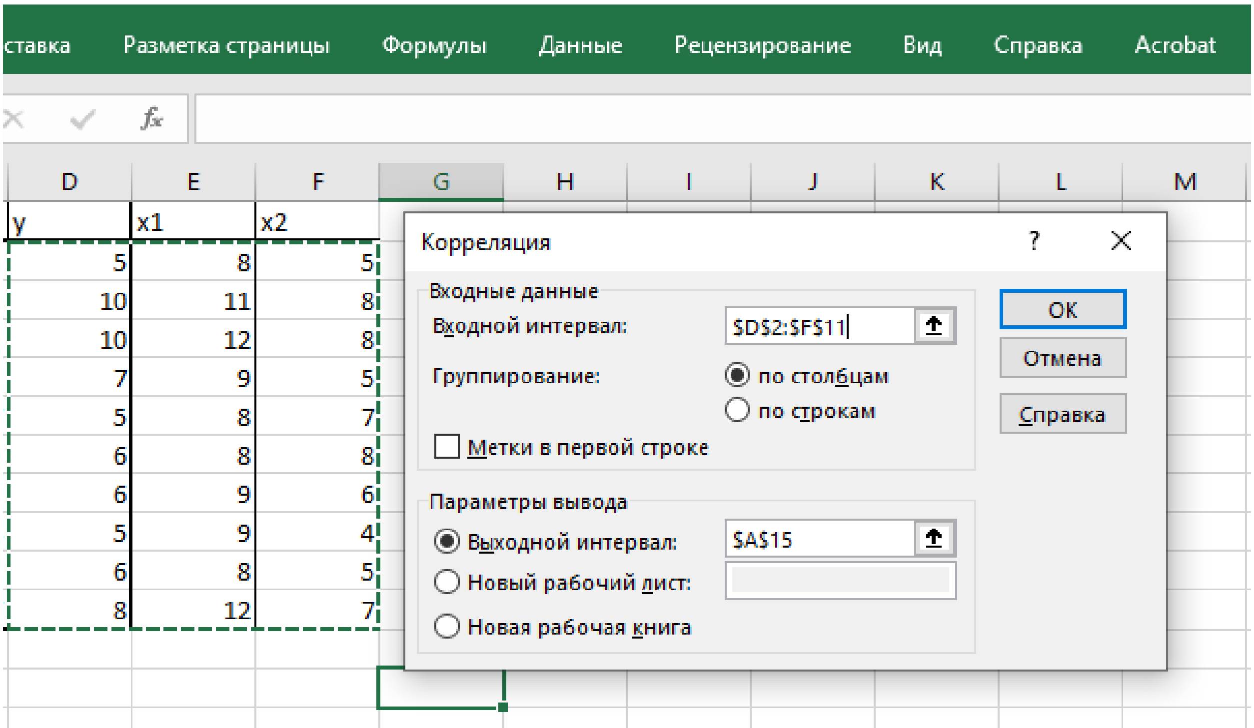


Рис. 20. Скриншот окна инструмента анализа «**Корреляция**».

Вид полученной корреляционной матрицы представлен на рис. 21:

	Столбец 1	*Столбец 2*	*Столбец 3*
Столбец 1	1		
Столбец 2	0,8661381	1	
Столбец 3	0,638764	0,48767536	1

Рис. 21. Скриншот корреляционной матрицы для рассматриваемых в задаче данных.

В полученной таблице:

«Столбец 1» - соответствует y;

«Столбец 2» - соответствует x_1;

«Столбец 3» - соответствует x_2;

Поскольку матрица линейной корреляции является симметрической, в ней достаточно указать главную диагональ и элементы под ней.

Если в окне инструмента анализа «**Корреляция**» установить флажок *Метки в первой строке*, выделив строку подписей над столбцами цифр, получим корреляционную матрицу вида, представленного на рис. 22.

	y	*x1*	*x2*
y	1		
x1	0,866138	1	
x2	0,638764	0,487675	1

Рис. 22. Скриншот корреляционной матрицы с установленным флажком *Метки в первой строке*.

Значения парных коэффициентов корреляции указывают на достаточно тесную связь сменной добычи угля на одного рабочего y с мощностью пласта x_1 и на умеренную связь с уровнем механизации работ x_2. В то же время межфакторная связь $r_{x_1x_2}$ не очень сильная ($r_{x_1x_2} = 0,49 < 0,7$), что говорит о том, что оба фактора являются информативными, т.е. и x_1, и x_2 необходимо включить в модель.

Независимо от формы связи показатель множественной корреляции может быть найден как индекс множественной корреляции по формуле 2.12:

$$R_{yx_1x_2\ldots x_m} = \sqrt{1 - \frac{\sigma^2_{\text{ост}}}{\sigma^2_y}},$$

где σ^2_y – общая дисперсия результативного признака;

$\sigma^2_{\text{ост}}$ – остаточная дисперсия.

Для определения остаточной дисперсии обратимся к группе показателей, объединенных в таблице ВЫВОД ОСТАТКА, данная таблица выводится, если в диалоговом окне режима **Регрессия** был задан параметр *Остатки*.

Скриншот части листа Excel с расположенной на нем таблицей ВЫВОД ОСТАТКА представлен на рисунке 23.

Столбец *Наблюдение* – содержит номера наблюдений;

Столбец *Предсказанное Y* – содержит значения $\hat{y}_i$, вычисленные по построенному уравнению регрессии;

Столбец *Остатки* – включает значения остатков $y_i - \hat{y}_i$;

ВЫВОД ОСТАТКА		
Наблюдение	*Предсказанное y*	*Остатки*
1	5,127340824	-0,127340824
2	8,790262172	1,209737828
3	9,644194757	0,355805243
4	5,981273408	1,018726592
5	5,861423221	-0,861423221
6	6,228464419	-0,228464419
7	6,348314607	-0,348314607
8	5,61423221	-0,61423221
9	5,127340824	0,872659176
10	9,277153558	-1,277153558

Рис. 23. Таблица ВЫВОД ОСТАТКА

Использование статистической функции Excel КВАДРОТКЛ() для расчета дисперсии уже было рассмотрено ранее. Математическая функция КОРЕНЬ() также была использована выше. Результат расчета дисперсии остатков представлен в 14 строке столбца *Остатки* таблицы на рис.24.

1	Исходные данные задачи				
2	№ п.п.	x_1	x_2	y	*Остатки*
3	1	8	5	5	-0,127340824
4	2	11	8	10	1,209737828
5	3	12	8	10	0,355805243
6	4	9	5	7	1,018726592
7	5	8	7	5	-0,861423221
8	6	8	8	6	-0,228464419
9	7	9	6	6	-0,348314607
10	8	9	4	5	-0,61423221
11	9	8	5	6	0,872659176
12	10	12	7	8	-1,277153558
13	Среднее	9,4	6,3	6,8	
14	σ^2	2,44	2,01	3,36	0,632958801
15	σ	1,562049935	1,417744688	1,833030278	

Рис. 24. Таблица с результатами расчета дисперсий.

$$R_{yx_1x_2\ldots x_m} = \sqrt{1-\frac{\sigma^2_{\text{ост}}}{\sigma^2_y}} = \sqrt{1-\frac{0,633}{3,36}} = 0,900899$$

Поскольку рассматриваемая регрессия линейная, определение совокупного коэффициента корреляции возможно через матрицу парных коэффициентов корреляции по формуле 2.16:

$$R_{yx_1x_2,\ldots,x_m} = \sqrt{1-\frac{\Delta r}{\Delta r_{11}}},$$

где

$$\Delta r = \begin{vmatrix} 1 & r_{yx_1} & r_{yx_2} & \ldots & r_{yx_m} \\ r_{yx_1} & 1 & r_{x_1x_2} & \ldots & r_{x_1x_m} \\ r_{yx_2} & r_{x_2x_1} & 1 & \ldots & r_{x_2x_m} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ r_{yx_m} & r_{x_mx_1} & r_{x_mx_2} & \ldots & 1 \end{vmatrix}$$

– определитель матрицы парных коэффициентов корреляции;

$$\Delta r_{11} = \begin{vmatrix} 1 & r_{x_1x_2} & \dots & r_{x_1x_m} \\ r_{x_2x_1} & 1 & \dots & r_{x_2x_m} \\ \dots & \dots & \dots & \dots \\ r_{x_mx_1} & r_{x_mx_2} & \dots & 1 \end{vmatrix}$$

– определитель матрицы межфакторной корреляции.

Для вычисления определителей матриц нужно воспользоваться математической функцией МОПРЕД(массив), которая возвращает определитель квадратной матрицы. Скриншот окна **Вставка функции** при обращении к функции МОПРЕД представлен на рис. 25.

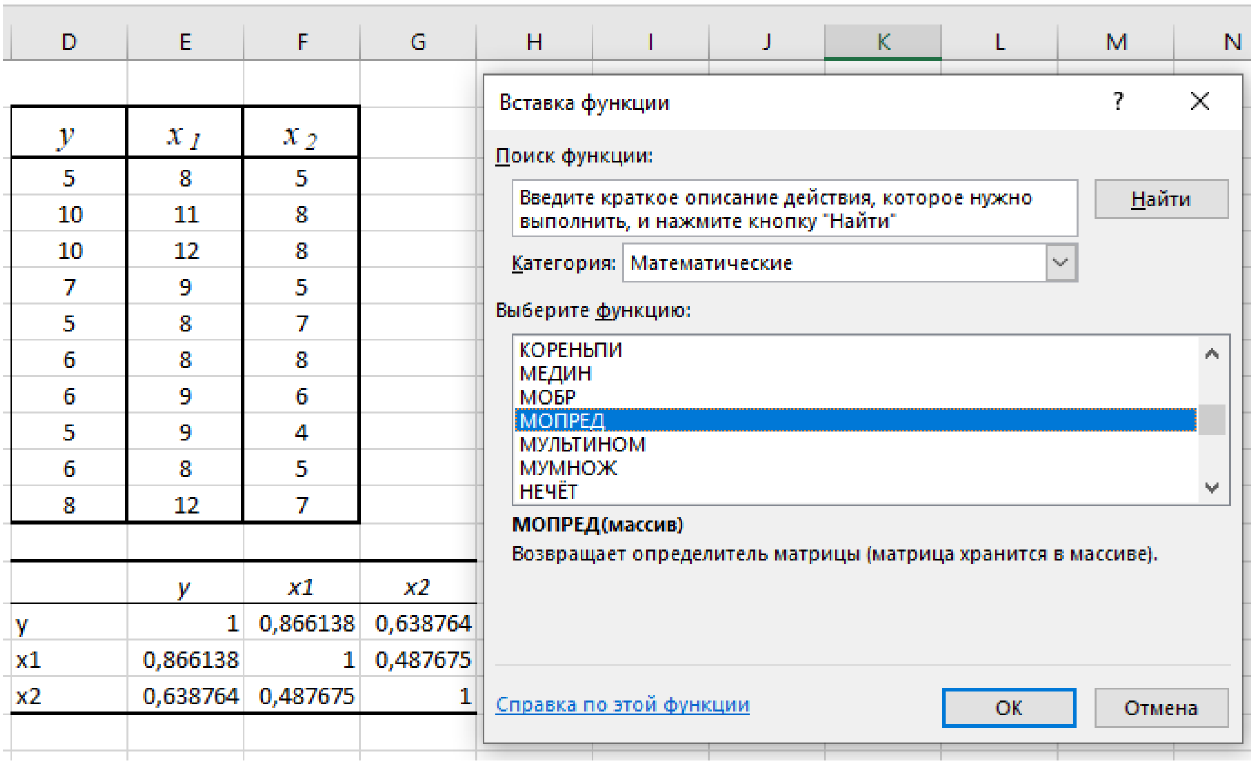


Рис. 25. Скриншот окна **Вставка функции** при обращении к функции МОПРЕД.

$$R_{yx_1x_2,...,x_m} = \sqrt{1-\frac{\Delta r}{\Delta r_{11}}} = \sqrt{1-\frac{0,143579}{0,762173}} = \sqrt{1-0,188381} = \sqrt{0,811619} =$$
$$= 0,900899$$

Определим совокупный коэффициент корреляции с использованием стандартизованных коэффициентов регрессии β_i по формуле 2.15. При линейной зависимости признаков формула индекса множественной корреляции (совокупного коэффициента корреляции) может быть представлена следующим выражением:

$$R_{yx_1x_2...x_m} = \sqrt{\sum \beta_i \cdot r_{yx_i}} ,$$

где β_i – стандартизованные коэффициенты регрессии;

r_{yx_i} – парные коэффициенты корреляции результата с каждым фактором.

Можно воспользоваться рассчитанными ранее коэффициентами β_i и r_{yx_i}.

$$R_{yx_1x_2...x_m} = \sqrt{\sum \beta_i \cdot r_{yx_i}} = \sqrt{0,72769 \cdot 0,866138+0,28389 \cdot 0,63876} =$$
$$= 0,900899$$

При всех вариантах расчета по разным формулам значения индекса детерминации совпадают между собой и со значением, полученным в результате работы функции **Регрессия**, которое приведено в таблице *Регрессионная статистика* на рис. 8.

Можно сказать, что 81,2% (коэффициент детерминации $R^2_{yx_1x_2} = 0,8116$) вариации результата объясняется вариацией представленных в уравнении регрессии признаков, что указывает на весьма тесную связь признаков с результатом.

Вычислим скорректированный коэффициент множественной детерминации по формуле 2.17а:

$$R = 1 - \left(1 - R^2\right) \cdot \frac{n-1}{n-m-1} = 1 - \left(1 - 0,8116\right) \cdot \frac{10-1}{10-2-1} = 0,758$$

Величина скорректированного коэффициента множественной детерминации указывает на заметную, но умеренную связь между результатом и признаками. Это можно объяснить также малым количеством наблюдений, использованных при решении задачи.

Рассчитаем частные коэффициенты корреляции по формуле 2.18а и определим по ним меру тесноты связи каждого фактора с результатом в чистом виде, чтобы иметь возможность ранжировать факторы по тесноте их связи с результатом на основе частных коэффициентов корреляции.

$$r_{yx_1 \cdot x_2} = \sqrt{1 - \frac{1 - R^2_{yx_1x_2}}{1 - r^2_{yx_2}}} = \sqrt{1 - \frac{1 - 0,8116}{1 - (0,6388)^2}} = 0.8257;$$

$$r_{yx_2 \cdot x_1} = \sqrt{1 - \frac{1 - R^2_{yx_1x_2}}{1 - r^2_{yx_1}}} = \sqrt{1 - \frac{1 - 0,8116}{1 - (0,4877)^2}} = 0,49587$$

Теперь найдем частные коэффициенты корреляции по формулам (2.19а):

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1x_2}}{\sqrt{\left(1 - r^2_{yx_2}\right) \cdot \left(1 - r^2_{x_1x_2}\right)}} = \frac{0,8661 - 0,6388 \cdot 0,4877}{\sqrt{\left(1 - 0,6388^2\right)\left(1 - 0,4877^2\right)}} = 0,8257$$

$$r_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1x_2}}{\sqrt{\left(1 - r^2_{yx_1}\right) \cdot \left(1 - r^2_{x_1x_2}\right)}} = \frac{0,6388 - 0,8661 \cdot 0,4877}{\sqrt{\left(1 - 0,8661^2\right)\left(1 - 0,4877^2\right)}} = 0,49587$$

Значения частных коэффициентов корреляции, рассчитанные по обеим формулам, позволяют сделать вывод, что фактор x_1 оказывает более сильное влияние на результат, чем признак x_2.

Надежность уравнения регрессии в целом и показателя связи была оценена с помощью F-критерия Фишера, табличное значение которого при пятипроцентном уровне значимости ($\alpha = 0{,}05$, $k_1 = 2$, $k_2 = 10 - 2 - 1 = 7$): $F_{\text{табл}} = 4{,}74$. Расчетное значение, приведенное в таблице на рис.7, получено при обращении к функции **Регрессия** $F_{\text{факт}} = 15{,}079$, что больше критического (табличного) в несколько раз, следовательно, уравнение регрессии признается статистически значимым.

Для того, чтобы оценить целесообразность включения фактора x_1 после фактора x_2 и наоборот, фактора x_2 после x_1, выполним расчет частных F-критериев Фишера по формулам (2.23а):

$$F_{x_1} = \frac{R^2_{yx_1x_2} - r^2_{yx_2}}{1 - R^2_{yx_1x_2}} \cdot (n-3) = \frac{0{,}8116 - 0{,}6388^2}{1 - 0{,}8116} \cdot 7 = 14{,}997;$$

$$F_{x_2} = \frac{R^2_{yx_1x_2} - r^2_{yx_1}}{1 - R^2_{yx_1x_2}} \cdot (n-3) = \frac{0{,}8116 - 0{,}8661^2}{1 - 0{,}8116} \cdot 7 = 2{,}28.$$

Табличное значение частного F-критерия при пятипроцентном уровне значимости ($\alpha = 0{,}05$, $k_1 = 1$, $k_2 = 10 - 2 - 1 = 7$): $F_{\text{табл}} = 5{,}59$. Так как $F_{x_1} = 14{,}997 > F_{\text{табл}} = 5{,}59$, то включение фактора x_1 в модель статистически оправдано и коэффициент чистой регрессии b_1 статистически значим, а поскольку $F_{x_2} = 2{,}28 < F_{\text{табл}} = 5{,}59$, то дополнительное включение фактора x_2, после того, как уже введен фактор x_1, нецелесообразно.

Область листа Excel с рассчитанными по формулам (2.18а) и (2.19а) значениями частных коэффициентов корреляции, а также рассчитанными по формулам (2.23а) значениями частных F-критериев Фишера представлены на рис. 26.

$\boxed{r_{yx_1 \cdot x_2} =}$	0,825699
$\boxed{r_{yx_2 \cdot x_1} =}$	0,495872
$\boxed{r_{yx_1 \cdot x_2} =}$	0,825699
$\boxed{r_{yx_2 \cdot x_1} =}$	0,495872
$F_{x_1} =$	14,9973
$F_{x_2} =$	2,282452

Рис. 26. Скриншот листа Excel со значениями частных коэффициентов корреляции и частных F-критериев Фишера.

Уравнение регрессии, включающее только один значимый аргумент x_1:

$$y = -2,754 + 1,016x_1 .$$

Скриншот результатов построения парной линейной регрессии с одним фактором x_1 представлен на рис. 27.

Уравнение в целом признается статистически значимым, однако нельзя исключить возможности улучшения модели путем проведения работы по спецификации модели. Это касается не только отбора факторов, но и выбора функции. Итогом работы должна стать модель, которая будет статистически значимой, будет иметь статистически значимые параметры, хорошо аппроксимировать исходные данные и удовлетворять предпосылкам метода наименьших квадратов, который мы использовали в данной задаче для оценки параметров модели.

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множеств	0,866138072
R-квадрат	0,75019516
Нормиро	0,718969555
Стандарт	1,024295039
Наблюде	10

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1	25,20656	25,20656	24,025	0,001191
Остаток	8	8,393443	1,04918		
Итого	9	33,6			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересеч	-2,754098361	1,975937	-1,39382	0,20087	-7,31062	1,80242
x1	1,016393443	0,207362	4,90153	0,001191	0,538215	1,494572

Рис. 27. Скриншот листа Excel с результатами работы функции Регрессия при построении регрессии, включающей только один значимый аргумент x_1.

5. Задания для самостоятельной работы

Задача 1.

По 20 предприятиям региона изучается зависимость выработки продукции на одного работника y (тыс. руб.) от ввода в действие новых основных фондов x_1 (% от стоимости фондов на конец года) и от удельного веса рабочих высокой квалификации в общей численности рабочих x_2 (%). Данные наблюдений для анализа приведены в таблице 5.1.

Таблица 5.1

Номер предприятия	y	x_1	x_2
1	**2**	**3**	**4**
1	7,0	3,9	10,0
2	7,0	3,9	14,0
3	7,0	3,7	15,0
4	7,0	4,0	16,0
5	7,0	3,8	17,0
6	7,0	4,8	19,0
7	8,0	5,4	19,0
8	8,0	4,4	20,0
9	8,0	5,3	20,0
10	10,0	6,8	20,0
11	9,0	6,0	21,0
12	11,0	6,4	22,0
13	9,0	6,8	22,0
14	11,0	7,2	25,0
15	12,0	8,0	28,0
16	12,0	8,2	29,0
17	12,0	8,1	30,0
18	12,0	8,5	31,0
19	14,0	9,6	32,0
20	14,0	9,0	36,0

Требуется:

1. Построить линейную модель множественной регрессии. Записать стандартизованное уравнение множественной регрессии. На основе стандартизованных коэффициентов регрессии и средних коэффициентов эластичности ранжировать факторы по степени их влияния на результат.

2. Найти коэффициенты парной, частной и множественной корреляции. Проанализировать их.

3. Найти скорректированный коэффициент множественной детерминации. Сравнить его с нескорректированным (общим) коэффициентом детерминации.

4. С помощью F-критерия Фишера оценить статистическую надежность уравнения регрессии и коэффициента детерминации $R^2_{yx_1x_2}$.

5. С помощью частных F-критериев Фишера оценить целесообразность включения в уравнение множественной регрессии фактора x_1 после x_2 и фактора x_2 после x_1.

6. Составить уравнение линейной парной регрессии, оставив лишь один значащий фактор.

Задача 2.

По 20 предприятиям региона изучается зависимость выработки продукции на одного работника y (тыс. руб.) от ввода в действие новых основных фондов x_1 (% от стоимости фондов на конец года) и от удельного веса рабочих высокой квалификации в общей численности рабочих x_2 (%). Данные наблюдений для анализа приведены в таблице 5.2.

Таблица 5.2

Номер предприятия	y	x_1	x_2
1	**2**	**3**	**4**
1	6	3,6	12
2	6	3,9	14
3	7	4,1	17
4	7	3,9	18
5	7	4,5	19
6	8	5,3	19
7	8	5,3	19
8	9	5,6	20
9	10	6,8	21
10	6	3,6	9
11	9	6,3	21
12	11	6,4	22
13	11	7,0	24
14	12	7,5	25
15	12	7,9	28
16	13	8,2	30
17	13	8,0	30
18	13	8,6	31
19	14	9,5	33
20	14	9,0	36

Требуется:

1. Построить линейную модель множественной регрессии. Записать стандартизованное уравнение множественной регрессии. На основе стандартизованных коэффициентов регрессии и средних коэффициентов эластичности ранжировать факторы по степени их влияния на результат.

2. Найти коэффициенты парной, частной и множественной корреляции. Проанализировать их.

3. Найти скорректированный коэффициент множественной детерминации. Сравнить его с нескорректированным (общим) коэффициентом детерминации.

4. С помощью F-критерия Фишера оценить статистическую надежность уравнения регрессии и коэффициента детерминации $R^2_{yx_1x_2}$.

5. С помощью частных F-критериев Фишера оценить целесообразность включения в уравнение множественной регрессии фактора x_1 после x_2 и фактора x_2 после x_1.

6. Составить уравнение линейной парной регрессии, оставив лишь один значащий фактор.

Задача 3.

По 20 предприятиям региона изучается зависимость выработки продукции на одного работника y (тыс. руб.) от ввода в действие новых основных фондов x_1 (% от стоимости фондов на конец года) и от удельного веса рабочих высокой квалификации в общей численности рабочих x_2 (%). Данные наблюдений для анализа приведены в таблице 5.3.

Таблица 5.3

Номер предприятия	y	x_1	x_2
1	**2**	**3**	**4**
1	6	3,6	12
2	7	3,9	15
3	7	4,1	17
4	7	4,2	18
5	8	4,5	19
6	8	5,3	19
7	9	5,3	20
8	9	5,6	20
9	10	6,0	21
10	6	3,5	10
11	10	6,3	21
12	11	6,4	22
13	11	7,0	23
14	12	7,5	25
15	12	7,9	28
16	13	8,2	30
17	13	8,4	31
18	14	8,6	31
19	14	9,5	35
20	15	10,0	36

Требуется:

1. Построить линейную модель множественной регрессии. Записать стандартизованное уравнение множественной регрессии. На основе стандартизованных коэффициентов регрессии и средних коэффициентов эластичности ранжировать факторы по степени их влияния на результат.

2. Найти коэффициенты парной, частной и множественной корреляции. Проанализировать их.

3. Найти скорректированный коэффициент множественной детерминации. Сравнить его с нескорректированным (общим) коэффициентом детерминации.

4. С помощью F-критерия Фишера оценить статистическую надежность уравнения регрессии и коэффициента детерминации $R^2_{yx_1x_2}$.

5. С помощью частных F-критериев Фишера оценить целесообразность включения в уравнение множественной регрессии фактора x_1 после x_2 и фактора x_2 после x_1.

6. Составить уравнение линейной парной регрессии, оставив лишь один значащий фактор.

Задача 4.

По 20 предприятиям региона изучается зависимость выработки продукции на одного работника y (тыс. руб.) от ввода в действие новых основных фондов x_1 (% от стоимости фондов на конец года) и от удельного веса рабочих высокой квалификации в общей численности рабочих x_2 (%). Данные наблюдений для анализа приведены в таблице 5.4.

Таблица 5.4

Номер предприятия	y	x_1	x_2
1	**2**	**3**	**4**
1	7	3,7	11
2	7	3,9	11
3	7	4,1	15
4	8	4,2	17
5	8	4,9	19
6	8	5,3	19
7	9	5,1	20
8	10	5,6	20
9	10	6,1	21
10	7	3,7	9
11	11	6,3	22
12	11	6,4	22
13	11	7,2	23
14	12	7,5	25
15	12	7,9	27
16	13	8,1	30
17	13	8,4	31
18	13	8,6	32
19	14	9,5	35
20	15	9,5	36

Требуется:

1. Построить линейную модель множественной регрессии. Записать стандартизованное уравнение множественной регрессии. На основе стандартизованных коэффициентов регрессии и средних коэффициентов эластичности ранжировать факторы по степени их влияния на результат.

2. Найти коэффициенты парной, частной и множественной корреляции. Проанализировать их.

3. Найти скорректированный коэффициент множественной детерминации. Сравнить его с нескорректированным (общим) коэффициентом детерминации.

4. С помощью F-критерия Фишера оценить статистическую надежность уравнения регрессии и коэффициента детерминации $R^2_{yx_1x_2}$.

5. С помощью частных F-критериев Фишера оценить целесообразность включения в уравнение множественной регрессии фактора x_1 после x_2 и фактора x_2 после x_1.

6. Составить уравнение линейной парной регрессии, оставив лишь один значащий фактор.

Задача 5.

По 20 предприятиям региона изучается зависимость выработки продукции на одного работника y (тыс. руб.) от ввода в действие новых основных фондов x_1 (% от стоимости фондов на конец года) и от удельного веса рабочих высокой квалификации в общей численности рабочих x_2 (%). Данные наблюдений для анализа приведены в таблице 5.5.

Таблица 5.5

Номер предприятия	y	x_1	x_2
1	**2**	**3**	**4**
1	7	3,6	10
2	7	3,9	12
3	7	4,1	17
4	8	4,2	18
5	8	4,5	19
6	9	5,3	19
7	9	5,5	20
8	10	5,6	21
9	10	6,1	21
10	7	3,5	9
11	10	6,3	22
12	10	6,5	22
13	11	7,2	24
14	12	7,5	25
15	12	7,9	27
16	13	8,2	30
17	13	8,4	31
18	14	8,6	33
19	14	9,5	35
20	15	9,6	36

Требуется:

1. Построить линейную модель множественной регрессии. Записать стандартизованное уравнение множественной регрессии. На основе стандартизованных коэффициентов регрессии и средних коэффициентов эластичности ранжировать факторы по степени их влияния на результат.

2. Найти коэффициенты парной, частной и множественной корреляции. Проанализировать их.

3. Найти скорректированный коэффициент множественной детерминации. Сравнить его с нескорректированным (общим) коэффициентом детерминации.

4. С помощью F-критерия Фишера оценить статистическую надежность уравнения регрессии и коэффициента детерминации $R^2_{yx_1x_2}$.

5. С помощью частных F-критериев Фишера оценить целесообразность включения в уравнение множественной регрессии фактора x_1 после x_2 и фактора x_2 после x_1.

6. Составить уравнение линейной парной регрессии, оставив лишь один значащий фактор.

Задача 6.

По 20 предприятиям региона изучается зависимость выработки продукции на одного работника y (тыс. руб.) от ввода в действие новых основных фондов x_1 (% от стоимости фондов на конец года) и от удельного веса рабочих высокой квалификации в общей численности рабочих x_2 (%). Данные наблюдений для анализа приведены в таблице 5.6.

Таблица 5.6

Номер предприятия	y	x_1	x_2
1	**2**	**3**	**4**
1	7	3,6	11
2	7	3,7	12
3	8	4,1	16
4	8	4,3	19
5	8	4,5	19
6	9	5,4	20
7	9	5,5	20
8	10	5,8	21
9	10	6,1	21
10	7	3,6	9
11	10	6,3	21
12	11	6,9	23
13	11	7,2	24
14	12	7,8	25
15	13	8,1	27
16	13	8,2	29
17	13	8,4	31
18	14	8,8	33
19	14	9,5	35
20	14	9,7	34

Требуется:

1. Построить линейную модель множественной регрессии. Записать стандартизованное уравнение множественной регрессии. На основе стандартизованных коэффициентов регрессии и средних коэффициентов эластичности ранжировать факторы по степени их влияния на результат.

2. Найти коэффициенты парной, частной и множественной корреляции. Проанализировать их.

3. Найти скорректированный коэффициент множественной детерминации. Сравнить его с нескорректированным (общим) коэффициентом детерминации.

4. С помощью F-критерия Фишера оценить статистическую надежность уравнения регрессии и коэффициента детерминации $R^2_{yx_1x_2}$.

5. С помощью частных F-критериев Фишера оценить целесообразность включения в уравнение множественной регрессии фактора x_1 после x_2 и фактора x_2 после x_1.

6. Составить уравнение линейной парной регрессии, оставив лишь один значащий фактор.

Задача 7.

По 20 предприятиям региона изучается зависимость выработки продукции на одного работника y (тыс. руб.) от ввода в действие новых основных фондов x_1 (% от стоимости фондов на конец года) и от удельного веса рабочих высокой квалификации в общей численности рабочих x_2 (%). Данные наблюдений для анализа приведены в таблице 5.7.

Таблица 5.7

Номер предприятия	y	x_1	x_2
1	**2**	**3**	**4**
1	7	3,6	10
2	7	3,8	14
3	7	4,2	15
4	8	4,3	18
5	8	4,7	19
6	9	5,4	19
7	9	5,6	20
8	10	5,9	20
9	10	6,1	21
10	7	3,5	9
11	10	6,3	21
12	10	6,8	22
13	11	7,2	24
14	12	7,9	25
15	12	8,1	26
16	13	8,3	29
17	13	8,4	31
18	13	8,8	32
19	14	9,6	35
20	14	9,7	36

Требуется:

1. Построить линейную модель множественной регрессии. Записать стандартизованное уравнение множественной регрессии. На основе стандартизованных коэффициентов регрессии и средних коэффициентов эластичности ранжировать факторы по степени их влияния на результат.

2. Найти коэффициенты парной, частной и множественной корреляции. Проанализировать их.

3. Найти скорректированный коэффициент множественной детерминации. Сравнить его с нескорректированным (общим) коэффициентом детерминации.

4. С помощью F-критерия Фишера оценить статистическую надежность уравнения регрессии и коэффициента детерминации $R^2_{yx_1x_2}$.

5. С помощью частных F-критериев Фишера оценить целесообразность включения в уравнение множественной регрессии фактора x_1 после x_2 и фактора x_2 после x_1.

6. Составить уравнение линейной парной регрессии, оставив лишь один значащий фактор.

Задача 8.

По 20 предприятиям региона изучается зависимость выработки продукции на одного работника y (тыс. руб.) от ввода в действие новых основных фондов x_1 (% от стоимости фондов на конец года) и от удельного веса рабочих высокой квалификации в общей численности рабочих x_2 (%). Данные наблюдений для анализа приведены в таблице 5.8.

Таблица 5.8

Номер предприятия	y	x_1	x_2
1	**2**	**3**	**4**
1	7	3,8	12
2	7	3,9	16
3	7	4,1	17
4	7	4,6	18
5	8	4,5	18
6	8	5,3	19
7	9	5,5	20
8	9	6,1	20
9	10	6,8	21
10	7	3,8	11
11	10	6,8	21
12	11	7,4	23
13	11	7,8	24
14	12	7,5	26
15	12	7,9	28
16	12	8,1	30
17	13	8,4	31
18	13	8,7	32
19	13	9,5	33
20	14	9,7	35

Требуется:

1. Построить линейную модель множественной регрессии. Записать стандартизованное уравнение множественной регрессии. На основе стандартизованных коэффициентов регрессии и средних коэффициентов эластичности ранжировать факторы по степени их влияния на результат.

2. Найти коэффициенты парной, частной и множественной корреляции. Проанализировать их.

3. Найти скорректированный коэффициент множественной детерминации. Сравнить его с нескорректированным (общим) коэффициентом детерминации.

4. С помощью F-критерия Фишера оценить статистическую надежность уравнения регрессии и коэффициента детерминации $R^2_{yx_1x_2}$.

5. С помощью частных F-критериев Фишера оценить целесообразность включения в уравнение множественной регрессии фактора x_1 после x_2 и фактора x_2 после x_1.

6. Составить уравнение линейной парной регрессии, оставив лишь один значащий фактор.

Задача 9.

По 20 предприятиям региона изучается зависимость выработки продукции на одного работника y (тыс. руб.) от ввода в действие новых основных фондов x_1 (% от стоимости фондов на конец года) и от удельного веса рабочих высокой квалификации в общей численности рабочих x_2 (%). Данные наблюдений для анализа приведены в таблице 5.9.

Таблица 5.9

Номер предприятия	y	x_1	x_2
1	**2**	**3**	**4**
1	7	4,1	14
2	7	4,3	16
3	7	4,1	17
4	8	4,6	17
5	8	4,7	18
6	9	5,3	20
7	9	5,5	20
8	11	6,9	21
9	10	6,8	21
10	7	3,8	9
11	11	7,1	22
12	11	7,5	23
13	12	7,8	25
14	12	7,6	27
15	12	7,9	29
16	13	8,1	30
17	13	8,5	32
18	14	8,7	32
19	14	9,6	33
20	15	9,8	36

Требуется:

1. Построить линейную модель множественной регрессии. Записать стандартизованное уравнение множественной регрессии. На основе стандартизованных коэффициентов регрессии и средних коэффициентов эластичности ранжировать факторы по степени их влияния на результат.

2. Найти коэффициенты парной, частной и множественной корреляции. Проанализировать их.

3. Найти скорректированный коэффициент множественной детерминации. Сравнить его с нескорректированным (общим) коэффициентом детерминации.

4. С помощью F-критерия Фишера оценить статистическую надежность уравнения регрессии и коэффициента детерминации $R^2_{yx_1x_2}$.

5. С помощью частных F-критериев Фишера оценить целесообразность включения в уравнение множественной регрессии фактора x_1 после x_2 и фактора x_2 после x_1.

6. Составить уравнение линейной парной регрессии, оставив лишь один значащий фактор.

Задача 10.

По 20 предприятиям региона изучается зависимость выработки продукции на одного работника y (тыс. руб.) от ввода в действие новых основных фондов x_1 (% от стоимости фондов на конец года) и от удельного веса рабочих высокой квалификации в общей численности рабочих x_2 (%). Данные наблюдений для анализа приведены в таблице 5.10.

Таблица 5.10

Номер предприятия	y	x_1	x_2
1	**2**	**3**	**4**
1	7	4,2	13
2	7	4,3	15
3	7	4,4	17
4	8	4,6	18
5	8	4,8	19
6	9	5,3	19
7	9	5,7	20
8	10	6,9	21
9	10	6,8	21
10	7	3,9	12
11	11	7,1	22
12	12	7,5	25
13	13	7,8	26
14	12	7,9	27
15	13	8,1	30
16	13	8,4	31
17	13	8,6	32
18	14	8,8	32
19	14	9,6	34
20	14	9,9	36

Требуется:

1. Построить линейную модель множественной регрессии. Записать стандартизованное уравнение множественной регрессии. На основе стандартизованных коэффициентов регрессии и средних коэффициентов эластичности ранжировать факторы по степени их влияния на результат.

2. Найти коэффициенты парной, частной и множественной корреляции. Проанализировать их.

3. Найти скорректированный коэффициент множественной детерминации. Сравнить его с нескорректированным (общим) коэффициентом детерминации.

4. С помощью F-критерия Фишера оценить статистическую надежность уравнения регрессии и коэффициента детерминации $R^2_{yx_1x_2}$.

5. С помощью частных F-критериев Фишера оценить целесообразность включения в уравнение множественной регрессии фактора x_1 после x_2 и фактора x_2 после x_1.

6. Составить уравнение линейной парной регрессии, оставив лишь один значащий фактор.

Задача 11.

По 20 предприятиям региона изучается зависимость выработки продукции на одного работника y (тыс. руб.) от ввода в действие новых основных фондов x_1 (% от стоимости фондов на конец года) и от удельного веса рабочих высокой квалификации в общей численности рабочих x_2 (%). Данные наблюдений для анализа приведены в таблице 5.11.

Таблица 5.11

Номер предприятия	y	x_1	x_2
1	**2**	**3**	**4**
1	7	4,1	14
2	7	4,3	16
3	7	4,4	17
4	7	4,5	18
5	8	4,8	19
6	8	5,3	20
7	8	5,6	20
8	9	6,7	21
9	10	6,9	22
10	7	3,6	12
11	10	7,2	23
12	11	7,6	25
13	12	7,8	26
14	11	7,9	28
15	12	8,2	30
16	12	8,4	31
17	12	8,6	32
18	13	8,8	32
19	13	9,2	33
20	14	9,6	34

Требуется:

1. Построить линейную модель множественной регрессии. Записать стандартизованное уравнение множественной регрессии. На основе стандартизованных коэффициентов регрессии и средних коэффициентов эластичности ранжировать факторы по степени их влияния на результат.

2. Найти коэффициенты парной, частной и множественной корреляции. Проанализировать их.

3. Найти скорректированный коэффициент множественной детерминации. Сравнить его с нескорректированным (общим) коэффициентом детерминации.

4. С помощью F-критерия Фишера оценить статистическую надежность уравнения регрессии и коэффициента детерминации $R^2_{yx_1x_2}$.

5. С помощью частных F-критериев Фишера оценить целесообразность включения в уравнение множественной регрессии фактора x_1 после x_2 и фактора x_2 после x_1.

6. Составить уравнение линейной парной регрессии, оставив лишь один значащий фактор.

Задача 12.

По совокупности страховых организаций изучается зависимость общей страховой суммы y (млн руб.) от количества заключенных договоров страхования x (тыс. ед.) и размера уставного капитала страховой организации z (млн руб.). Данные наблюдений для анализа приведены в таблице 5.12.

Таблица 5.12

Номер организации	x	z	y
1	**2**	**3**	**4**
1	20	390	51,3
2	22	380	57,2
3	23	380	53,9
4	25	410	63,2
5	27	420	61,1
6	28	420	67,8
7	30	550	73,5
8	32	520	80,0
9	34	410	68,1
10	35	450	79,0
11	37	470	75,6
12	44	460	91,4
13	45	480	85,1
14	46	520	98,2
15	49	510	91,6
16	52	450	101,1
17	57	460	96,9
18	59	450	110,2
19	64	500	107,4
20	66	550	126,3
21	68	590	118,1
22	70	800	149,0

Требуется:

1. Построить линейную модель множественной регрессии. Записать стандартизованное уравнение множественной регрессии. На основе стандартизованных коэффициентов регрессии и средних коэффициентов эластичности ранжировать факторы по степени их влияния на результат.

2. Найти коэффициенты парной, частной и множественной корреляции. Проанализировать их.

3. Найти скорректированный коэффициент множественной детерминации. Сравнить его с нескорректированным (общим) коэффициентом детерминации.

4. С помощью F-критерия Фишера оценить статистическую надежность уравнения регрессии и коэффициента детерминации $R^2_{yx_1x_2}$.

5. С помощью частных F-критериев Фишера оценить целесообразность включения в уравнение множественной регрессии фактора x_1 после x_2 и фактора x_2 после x_1.

6. Составить уравнение линейной парной регрессии, оставив лишь один значащий фактор.

Рекомендуемая литература

1. Методы эконометрики: Учебник / С.А. Айвазян; Московская школа экономики МГУ им. М.В. Ломоносова (МШЭ). - М.: Магистр: ИНФРА-М, 2010. - 512 с.: Режим доступа: http://znanium.com/bookread2.php?book=196548
2. Эконометрика и эконометрическое моделирование : учебник / Л.О. Бабешко, М.Г. Бич, И.В. Орлова. - М. : Вузовский учебник : ИНФРА-М, 2018. - 385 с. : ил. — (Высшее образование: Бакалавриат). Режим доступа: http://znanium.com/bookread2.php?book=
3. Эконометрика / Уткин В.Б., - 2-е изд. - М.:Дашков и К, 2017. - 564 с.: ISBN 978-5-394-02145-9 Режим доступа: http://znanium.com/bookread2.php?book=415317
4. Артамонов, Н. В. Введение в эконометрику: Учебник / Артамонов Н.В., - 2-е изд., испр. и доп. - Москва :МЦНМО, 2014. - 222 с.: ISBN 978-5-4439-2010-8. - Текст : электронный. - URL: https://znanium.com/catalog/product/969269. – Режим доступа: по подписке.
5. Эконометрика: учеб. / под ред. И.И. Елисеевой. – М.: Проспект. 2009. – 288с.
6. Эконометрика: учеб. / И.И. Елисеева, С.В. Курышева, Т.В. Костеева и др.; под ред. И.И. Елисеевой. – 2-е изд., перераб. и доп. - М.: Финансы и статистика, 2006. – 576 с.: ил.
7. Кремер Н.Ш. Эконометрика: учебник для студентов вузов / Н.Ш. Кремер, Б.А. Путко; под ред. Н.Ш. Кремера. — 3-е изд., перераб. и доп. — М.: ЮНИТИ-ДАНА, 2010. — 328 с. — (Серия «Золотой фонд российских учебников»).
8. Колпаков, В. Ф. Экономико-математическое и эконометрическое моделирование: компьютерный практикум: учеб. пособие / В.Ф. Колпаков. - Москва : ИНФРА-М, 2018. – 396 с. - (Высшее образование: Бакалавриат). - www.dx.doi.org/10.12737/24417. - ISBN 978-5-16-010967-1. - Текст : электронный. - URL: https://znanium.com/catalog/product/975797. – Режим доступа: по подписке.
9. Бородич, С. А. Эконометрика. Практикум : учебное пособие / С.А. Бородич. — Москва : ИНФРА-М, 2021. — 329 с. : ил. — (Высшее образование: Бакалавриат). - ISBN 978-5-16-009429-8. - Текст : электронный. - URL: https://znanium.com/catalog/product/1228789. – Режим доступа: по подписке.
10. Орлова, И. В. Эконометрика. Компьютерный практикум для студентов третьего курса, обучающихся по специальностям 080105.65 "Финансы и кредит", 080109.65 "Бухгалтерский учет, анализ и аудит" / И. В.

Орлова, Е. С. Филонова, А. В. Агеев. - Москва : ВЗФЭИ, 2011. - 96 с. - Текст : электронный. - URL: https://znanium.com/catalog/product/453458. – Режим доступа: по подписке.

11. Хайяши, Ф. Эконометрика / Ф. Хайяши ; пер. с англ. под науч. ред. В.П. Носко. — Москва : Издательский дом «Дело» РАНХиГС, 2017. — 728 с. — (Академический учебник). - ISBN 978-5-7749-1197-4. - Текст : электронный. - URL: https://znanium.com/catalog/product/1043302 (дата обращения: 11.01.2023). – Режим доступа: по подписке.
12. Экономико-математические методы в примерах и задачах : учебное пособие / И.В. Орлова, Н.В. Концевая, Е.Н. Горбатенко, В.А. Большаков ; под ред. А.Н. Гармаша. — Москва : Вузовский учебник : ИНФРА-М, 2021. — 416 с. + Доп. материалы [Электронный ресурс]. - ISBN 978-5-9558-0322-7. - Текст : электронный. - URL: https://znanium.com/catalog/product/1659052 (дата обращения: 11.01.2023). – Режим доступа: по подписке.
13. Эконометрика: краткий курс лекций для обучающихся направления подготовки «Менеджмент» / Сост.: В. А. Шибайкин // ФГБОУ ВО Саратовский ГАУ. – Саратов, 2017. – с.60
14. Доугерти К. Введение в эконометрику: Учебник. 3-е изд. / Пер. с англ. – М.: ИНФРА-М, 2009. – XIV, 465 с/ - (Университетский учебник).
15. Кузнецова О.А. Эконометрика: практикум / О.А. Кузнецова, О.Н. Мазурмович. – Самара:Изд-во Самарского университета, 2019. – 72 с.
16. Максимова Т.Г., Попова И.Н. Эконометрика: учебно-методическое пособие / Т.Г. Максимова, И.Н. Попова. – СПб.: Университет ИТМО, 2018. – 70 с.

Приложение 1

1.1. Таблица значений F-критерия Фишера при уровне значимости $\alpha = 0{,}05$

k_2 \ k_1	1	2	3	4	5	6	8	12	24	∞
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
1	161,5	199,5	215,7	224,6	230,2	233,9	238,9	243,9	249,0	254,3
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,45	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,64	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,77	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	3,84	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,12	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,07	2,90	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,91	2,74	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,79	2,61	2,40
12	4,75	3,88	3,49	3,26	3,11	3,00	2,85	2,69	2,50	2,30
13	4,67	3,80	3,41	3,18	3,02	2,92	2,77	2,60	2,42	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,53	2,35	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,48	2,29	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,42	2,24	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,38	2,19	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,34	2,15	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,31	2,11	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,28	2,08	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,42	2,25	2,05	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2,23	2,03	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,38	2,20	2,00	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,36	2,18	1,98	1,73
25	4,24	3,38	2,99	2,76	2,60	2,49	2,34	2,16	1,96	1,71
26	4,22	3,37	2,98	2,74	2,59	2,47	2,32	2,15	1,95	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,30	2,13	1,93	1,67
28	4,20	3,34	2,95	2,71	2,56	2,44	2,29	2,12	1,91	1,65
29	4,18	3,33	2,93	2,70	2,54	2,43	2,28	2,10	1,90	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,09	1,89	1,62
35	4,12	3,26	2,87	2,64	2,48	2,37	2,22	2,04	1,83	1,57
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,00	1,79	1,51
45	4,06	3,21	2,81	2,58	2,42	2,31	2,15	1,97	1,76	1,48

1	2	3	4	5	6	7	8	9	10	11
50	4,03	3,18	2,79	2,56	2,40	2,29	2,13	1,95	1,74	1,44
60	4,00	3,15	2,76	2,52	2,37	2,25	2,10	1,92	1,70	1,39
70	3,98	3,13	2,74	2,50	2,35	2,23	2,07	1,89	1,67	1,35
80	3,96	3,11	2,72	2,49	2,33	2,21	2,06	1,88	1,65	1,31
90	3,95	3,10	2,71	2,47	2,32	2,20	2,04	1,86	1,64	1,28
100	3,94	3,09	2,70	2,46	2,30	2,19	2,03	1,85	1,63	1,26
125	3,92	3,07	2,68	2,44	2,29	2,17	2,01	1,83	1,60	1,21
150	3,90	3,06	2,66	2,43	2,27	2,16	2,00	1,82	1,59	1,18
200	3,89	3,04	2,65	2,42	2,26	2,14	1,98	1,80	1,57	1,14
300	3,87	3,03	2,64	2,41	2,25	2,13	1,97	1,79	1,55	1,10
400	3,86	3,02	2,63	2,40	2,24	2,12	1,96	1,78	1,54	1,07
500	3,86	3,01	2,62	2,39	2,23	2,11	1,96	1,77	1,54	1,06
1000	3,85	3,00	2,61	2,38	2,22	2,10	1,95	1,76	1,53	1,03
∞	3,84	2,99	2,60	2,37	2,21	2,09	1,94	1,75	1,52	1

1.2. Критические значения t-критерия Стьюдента при уровне значимости 0,10, 0,05, 0,01 (двухсторонний)

Число степеней свободы d.f.	α			Число степеней свободы d.f.	α		
	00,10	0,05	0,01		00,10	0,05	0,01
1	6,3138	12,706	63,657	18	1,7341	2,1009	2,8784
2	2,9200	4,3027	9,9248	19	1,7291	2,0930	2,8609
3	2,3534	3,1825	5,8409	20	1,7247	2,0860	2,8453
4	2,1318	2,7764	4,5041	21	1,7207	2,0796	2,8314
5	2,0150	2,5706	4,0321	22	1,7171	2,0739	2,8188
6	1,9432	2,4469	3,7074	23	1,7139	2,0687	2,8073
7	1,8946	2,3646	3,4995	24	1,7109	2,0639	2,7969
8	1,8595	2,3060	3,3554	25	1,7081	2,0595	2,7874
9	1,8331	2,2622	3,2498	26	1,7056	2,0555	2,7787
10	1,8125	2,2281	3,1693	27	1,7033	2,0518	2,7707
11	1,7959	2,2010	3,1058	28	1,7011	2,0484	2,7633
12	1,7823	2,1788	3,0545	29	1,6991	2,0452	2,7564
13	1,7709	2,1604	3,0123	30	1,6973	2,0423	2,7500
14	1,7613	2,1448	2,9768	40	1,6839	2,0211	2,7045
15	1,7530	2,1315	2,9467	60	1,6707	2,0003	2,6603
16	1,7459	2,1199	2,9208	120	1,6577	1,9799	2,6174
17	1,7396	2,1098	2,8982	∞	1,6449	1,9600	2,5758

СОДЕРЖАНИЕ

Введение 3
1. Общие методические указания для решения задач по дисциплине «Эконометрика» 4
2. Множественная регрессия и корреляция 6
2.1. Экономические процессы, описываемые с помощью уравнений множественной регрессии 6
2.2. Спецификация модели. Отбор факторов при построении уравнения множественной регрессии 6
2.3. Метод наименьших квадратов (МНК). Свойства оценок на основе МНК 12
2.4. Проверка существенности факторов и показатели качества регрессии 17
3. Контрольные вопросы 28
4. Построение модели множественной регрессии. 29
5. Задания для самостоятельной работы 62
Рекомендуемая литература 86
Приложение 1. Математико-статистические таблицы 88

Учебное издание

А.А. Чалганова

Построение множественной регрессии и оценка качества модели с использованием табличного процессора Excel

Публикуется в авторской редакции.

Подписано к публикации 30.12.2022. Формат 60×90 1/16.
Гарнитура Times New Roman. Усл. печ. л. 5,625. Заказ № 1344.
РГГМУ, 192007, Санкт-Петербург, Воронежская ул., д. 79.