Министерство науки и высшего образования Российской Федерации
Читинский институт (филиал)
Байкальского государственного университета

Л. И. Трухина

ПРАКТИКУМ ПО ЭКОНОМЕТРИКЕ: ПАРНАЯ И МНОЖЕСТВЕННАЯ РЕГРЕССИЯ

Учебное пособие

Казань
Издательство «Бук»
2023

УДК 330.43(075.8)
ББК 65в631я73
Т80

Рецензенты:
Пешков Николай Валерьевич, кандидат физико-математических наук, кафедра математики и прикладной механики (Забайкальский институт железнодорожного транспорта — филиал Иркутского государственного университета путей сообщения)
Черхарова Наталья Ивановна, кандидат технических наук, доцент, кафедра информационных образовательных технологий (Кубанский государственный университет)

Трухина, Людмила Ивановна.
Т80 Практикум по эконометрике: парная и множественная регрессия : учебное пособие / Л. И. Трухина; М-во науки и высшего образования Рос. Федерации, Читинский ин-т (фил.) Байкальского гос. ун-та. — Казань : Бук, 2023. — 88 с. — Текст : непосредственный.

ISBN 978-5-907753-41-9.

Учебное пособие содержит краткое изложение теоретического материала по двум разделам эконометрики: «Парная регрессия и корреляция», «Множественная регрессия и корреляция»; примеры решения задач, в том числе с использованием табличного процессора MS Excel; задачи для самостоятельного решения, а также варианты заданий для выполнения домашних работ. Материалы практикума обеспечивают методическую поддержку практических занятий и позволяют приобрести навык решения задач базового уровня по курсу эконометрики.

Пособие предназначается для обучающихся экономических специальностей высших учебных заведений.

УДК 330.43(075.8)
ББК 65в631я73

ISBN 978-5-907753-41-9

Предисловие

Пособие ориентировано на начальный курс эконометрики. Изучение этой дисциплины предполагает приобретение студентами опыта построения эконометрических моделей, принятия решения о спецификации модели, выбора метода оценки параметров модели, интерпретации результатов, получения прогнозных оценок. В связи с этим курс эконометрики обязательно включает решение задач.

Построение эконометрических моделей требует большого объема вычислений, которые могут быть выполнены с помощью вычислительной техники и специального программного обеспечения. Одним из наиболее популярных и удобных инструментов для проведения эконометрического анализа является табличный процессор Excel. В учебном пособии подробно описывается методика решения задач регрессионного анализа с использованием Excel. Это позволяет студентам освоить основные принципы работы с данными и научиться строить эконометрические модели на практике.

Эконометрический анализ с использованием Excel имеет ряд преимуществ. Во-первых, Excel является доступным программным обеспечением, которое можно установить на любой компьютер. Во-вторых, Excel обладает широким функционалом для работы с табличными данными, что позволяет проводить сложные вычисления и анализировать большие объемы информации.

Практикум охватывает две темы курса: парную и множественную регрессию. Разделы имеют идентичную структуру:

- краткий теоретический материал;
- примеры решения типовых задач;
- задачи для самостоятельного решения;
- домашнее задание.

В конце пособия находятся основные статистико-математические таблицы, необходимые для решения задач.

Данное пособие поможет студентам изучить основы работы с данными, научиться строить и анализировать эконометрические модели, а также получить навыки, которые будут полезны в их будущей профессиональной деятельности.

ПАРНАЯ РЕГРЕССИЯ

Понятие парной регрессии

Парной регрессией называется модель, выражающая зависимость среднего значения зависимой переменной y от одной независимой переменной x

$$\hat{y} = f(x) \text{ или } y = f(x) + \varepsilon,$$

где y – зависимая переменная (результативный признак); x – независимая, объясняющая переменная (признак-фактор). $\hat{y}$ − теоретическое значение результативного признака, найденное, исходя из уравнения регрессии; ε – случайная величина, характеризующая отклонения реального значения результативного признака от теоретического, найденного по уравнению регрессии.

Парная регрессия применяется, если имеется доминирующий фактор, обуславливающий большую долю изменения изучаемой объясняемой переменной, который и используется в качестве объясняющей переменной.

Постановка задачи

По имеющимся данным n наблюдений за совместным изменением двух переменных показателей x и y $\{(x_i, y_i), i = 1,2, ..., n\}$ необходимо определить аналитическую зависимость $\hat{y} = f(x)$, наилучшим образом описывающую данные наблюдений.

Результаты наблюдений удобно представлять в виде таблицы (табл. 1)

Таблица 1

№	x	y
1	x_1	y_1
2	x_2	y_2
…	…	…
n	x_n	y_n

Каждая строка таблицы представляет собой результат одного наблюдения (x_i, y_i).

Линейная модель парной регрессии и корреляции

Линейная парная регрессия описывается уравнением:

$$\hat{y} = a + b \cdot x \text{ или } y = a + b \cdot x + \varepsilon,$$

согласно которому изменение Δy переменной y прямо пропорционально изменению Δx переменной x ($\Delta y = b \cdot \Delta x$).

Линейная регрессия находит широкое применение в эконометрике ввиду четкой экономической интерпретации ее параметров.

Оценка параметров линейной парной регрессии

Построение линейной регрессии сводится к оценке ее параметров a и b. Классический подход к оцениванию параметров линейной регрессии основан на методе наименьших квадратов (МНК).

Согласно МНК, выбираются такие значения параметров a и b, при которых сумма квадратов отклонений фактических значений результативного признака y_i от теоретических значений $\hat{y}_i = f(x_i)$ (при тех же значениях фактора x_i) минимальна, т. е.

$$S = \sum (y_i - \hat{y}_i)^2 \to min.$$

Выполняя соответствующие вычисления и некоторые преобразования, получается система нормальных уравнений метода наименьших квадратов относительно a и b:

$$\begin{cases} n \cdot a + b \cdot \sum x_i = \sum y_i ; \\ a \cdot \sum x_i + b \cdot \sum x_i^2 = \sum x_i \cdot y_i . \end{cases}$$

Используя соотношения

$$n\overline{x} = \sum x_i, \quad n\overline{y} = \sum y_i , \quad n\overline{x^2} = \sum x_i^2, \quad n\overline{xy} = \sum x_i y_i,$$

получим

$$\begin{cases} a + b \cdot \overline{x} = \overline{y}; \\ a \cdot \overline{x} + b \cdot \overline{x^2} = \overline{x \cdot y}. \end{cases}$$

Из последней системы можно получить готовые формулы для определения параметров a и b:

$$b = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2},$$

$$a = \overline{y} - b \cdot \overline{x}.$$

Формулу для параметра b можно представить следующим образом

$$b = \frac{cov(x,y)}{\sigma_x^2},$$

где $cov(x,y) = \overline{xy} - \bar{y} \cdot \bar{x}$ – ковариация признаков x и y,

$\sigma_x^2 = \overline{x^2} - \overline{x}^2$ – дисперсия признака x.

Оценка тесноты связи

Для оценки тесноты связи изучаемых явлений используют линейный коэффициент корреляции r_{xy}, который можно рассчитать по следующим формулам:

$$r_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n \cdot \sigma_x \cdot \sigma_y} = \frac{cov(x,y)}{\sigma_x \cdot \sigma_y} = \frac{\overline{x \cdot y} - \bar{y} \cdot \bar{x}}{\sigma_x \cdot \sigma_y},$$

$\sigma_x = \sqrt{\overline{x^2} - \overline{x}^2}$ – среднеквадратическое отклонение x,

$\sigma_y = \sqrt{\overline{y^2} - \overline{y}^2}$ – среднеквадратическое отклонение y.

Линейный коэффициент корреляции находится в пределах:

$$-1 \le r_{xy} \le 1.$$

Чем ближе абсолютное значение r_{xy} к единице, тем сильнее линейная связь между факторами (при $r_{xy} = \pm 1$ имеем строгую функциональную зависимость).

Для качественной оценки тесноты связи можно использовать следующую классификацию:

$0{,}1 \le |r_{xy}| \le 0{,}3$ – очень слабая связь;

$0{,}3 \le |r_{xy}| \le 0{,}5$ – слабая связь;

$0{,}5 \le |r_{xy}| \le 0{,}7$ – умеренная связь;

$0{,}7 \le |r_{xy}| \le 0{,}9$ – тесная связь;

$0{,}9 \le |r_{xy}| \le 0{,}99$ – очень тесная.

Знак коэффициента корреляции указывает направление связи. Если $r_{xy} > 0$, то связь прямая; если $r_{xy} < 0$, то связь обратная.

Коэффициент линейной парной корреляции может быть определен через коэффициент регрессии b:

$$r_{xy} = b\frac{\sigma_x}{\sigma_y}.$$

Оценка качества построенной модели

Для оценки качества подбора линейной функции рассчитывается квадрат линейного коэффициента корреляции r_{xy}^2, называемый коэффициентом детерминации. Коэффициент детерминации характеризует долю дисперсии результативного признака *y*, объясняемую регрессией, в общей дисперсии результативного признака:

$$r_{xy}^2 = \frac{\sigma_{факт}^2}{\sigma_y^2} = 1 - \frac{\sigma_{ост}^2}{\sigma_y^2},$$

где

$\sigma_y^2 = \frac{1}{n}\sum_i (y_i - \bar{y})^2$ – общая дисперсия результативного признака *y*,

$\sigma_{факт}^2 = \frac{1}{n}\sum_i (\hat{y}_i - \bar{y})^2$ – факторная, объяснённая дисперсия,

$\sigma_{ост}^2 = \frac{1}{n}\sum_i (y_i - \hat{y}_i)^2$– остаточная дисперсия.

Соответственно величина $1 - r_{xy}^2$ характеризует долю дисперсии *y*, вызванную влиянием остальных, не учтенных в модели, факторов.

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

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

Средняя ошибка аппроксимации не должна превышать 8 – 10%.

Оценка значимости уравнения регрессии в целом и его параметров

Оценка значимости уравнения регрессии в целом производится на основе *F-критерия Фишера*, которому предшествует дисперсионный анализ.

Согласно основной идее дисперсионного анализа, общая сумма квадратов отклонений переменной *y* от среднего значения $\bar{y}$ раскладывается на две части – «объясненную» и «необъясненную»:

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

где $\sum_i (y_i - \bar{y})^2$ – общая сумма квадратов отклонений;

$\sum_i (\hat{y}_i - \bar{y})^2$ – сумма квадратов отклонений, объясненная регрессией (или факторная сумма квадратов отклонений);

$\sum_i (y_i - \hat{y}_i)^2$ – остаточная сумма квадратов отклонений, характеризующая влияние неучтенных в модели факторов.

Схема дисперсионного анализа имеет вид, представленный в таблице 2 (*n* – число наблюдений, *m* – число параметров при переменной *x*).

Таблица 2

Компоненты дисперсии	Сумма квадратов	Число степеней свободы	Дисперсия на одну степень свободы
Общая	$\sum_i (y_i - \bar{y})^2$	$n - 1$	$S^2_{общ} = \frac{\sum_i (y_i - \bar{y})^2}{n-1}$
Факторная	$\sum_i (\hat{y}_i - \bar{y})^2$	m	$S^2_{факт} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{m}$
Остаточная	$\sum_i (y_i - \hat{y}_i)^2$	$n - m - 1$	$S^2_{ост} = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-m-1}$

Определение дисперсии на одну степень свободы приводит дисперсии к сравнимому виду. Сопоставляя факторную и остаточную дисперсии в расчете на одну степень свободы, получим величину *F*-критерия Фишера:

$$F = \frac{S^2_{факт}}{S^2_{ост}} = \frac{\frac{\sum_i (\hat{y}_i - \bar{y})^2}{m}}{\frac{\sum_i (y_i - \hat{y}_i)^2}{n-m-1}} = \frac{r^2_{xy}}{1 - r^2_{xy}} \frac{n-m-1}{m}.$$

Проверяется гипотеза H_0 о статистической незначимости уравнения регрессии. Фактическое значение *F*-критерия Фишера сравнивается с табличным значением $F_{\text{таб}}(\alpha; k_1; k_2)$ 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 гипотеза H_0 отклоняется и признается статистическая значимость уравнения в целом.

Уровень значимости α – вероятность отвергнуть правильную гипотезу при условии, что она верна. Обычно величина α принимается равной 0,05 или 0,01.

Для парной линейной регрессии $m = 1$, поэтому

$$F = \frac{r^2_{xy}}{1 - r^2_{xy}}(n-2).$$

Для оценки статистической значимости отдельных параметров уравнения и парного линейного коэффициента корреляции рассчитываются *t-критерии Стьюдента*. Выдвигается гипотеза H_0 о случайной природе показателей, т.е. о незначимом их отличии от нуля.

Рассчитываются фактические значения *t*-критерия путём сопоставления их значений с величиной стандартной ошибки:

$$t_b = \frac{b}{m_b};\ t_a = \frac{a}{m_a};\ t_r = \frac{r}{m_r}.$$

С этой целью по каждому из параметров определяется его стандартная ошибка.

Стандартные ошибки коэффициента регрессии b, параметра a и коэффициента корреляции r_{xy} определяются по формулам:

$$m_b = \sqrt{\frac{S_{ост}^2}{\sum(x-\bar{x})^2}} = \frac{S_{ост}}{\sigma_x\sqrt{n}}.$$

$$m_a = \sqrt{S_{ост}^2 \frac{\sum x^2}{n\sum(x-\bar{x})^2}} = S_{ост}\frac{\sqrt{\sum x^2}}{\sigma_x \cdot n}.$$

$$m_r = \sqrt{\frac{1-r_{xy}^2}{n-2}}.$$

Сравнивая фактические и табличные значения t-статистики при определенном уровне значимости α и числе степеней свободы (n – 2) принимаем или отвергаем гипотезу H_0. Если $t_{\text{табл}} < t_{\text{факт}}$, то H_0 отклоняется, т.е. параметр (a или b) или r_{xy} не случайно отличаются от нуля и сформировался под влиянием систематически действующего фактора x. Если $t_{\text{табл}} > t_{\text{факт}}$, то H_0 принимается и признается случайная природа формирования a, b или r_{xy}.

Существует связь между *t*-критерием Стьюдента и *F*-критерием Фишера:

$$t_r^2 = t_b^2 = F.$$

Таким образом, проверка гипотез о значимости коэффициента регрессии и корреляции равносильна проверке гипотезы о существенности линейного уравнения регрессии.

Рассчитанные значения показателей a, b, и r являются приближенными, полученными на основе имеющихся выборочных данных. Для того, чтобы оценить, насколько точные значения показателей могут отличаться от рассчитанных, строят доверительные интервалы.

Доверительный интервал для коэффициента регрессии определяется как

$$b - t_{\text{таб}} \cdot m_b \le b^* \le b + t_{\text{таб}} \cdot m_b.$$

Аналогично определяется доверительный интервал для параметра a

$$a - t_{\text{таб}} \cdot m_a \le a^* \le a + t_{\text{таб}} \cdot m_a.$$

Если в границы доверительного интервала попадает ноль, т.е. нижняя граница отрицательна, а верхняя положительна, то оцениваемый параметр принимается нулевым, т.к. он не может одновременно принимать и положительное и отрицательное значения.

Точечный и интервальный прогноз по уравнению линейной регрессии

Прогнозное значение $\hat{y}_p$ определяется путем подстановки в уравнение регрессии $\hat{y} = a + b \cdot x$ соответствующего прогнозного значения x_p. Получается точечный прогноз, который дополняется расчётом стандартной ошибки прогноза $m_{\hat{y}_p}$

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

и построением доверительного интервала прогноза

$$\hat{y}_p - t_{\text{таб}} \cdot m_{\hat{y}_p} \le \hat{y}_p^* \le \hat{y}_p + t_{\text{таб}} \cdot m_{\hat{y}_p}.$$

Доверительны интервал всегда определяется с заданной вероятностью, соответствующей принятому уровню значимости α.

Нелинейные модели парной регрессии и корреляции

Многие экономические зависимости не являются линейными по своей сути, и поэтому их моделирование линейными уравнениями регрессии, безусловно, не даст положительного результата. Поэтому они должны выражаться с помощью соответствующих нелинейных функций.

Различают два класса нелинейных регрессий:

1. Регрессии, *нелинейные относительно включенных в анализ объясняющих переменных, но линейные по оцениваемым параметрам*, например

- полиномы различных степеней: $\hat{y} = a + b \cdot x + c \cdot x^2$,

$$\hat{y} = a + b \cdot x + c \cdot x^2 + d \cdot x^3;$$

- гипербола: $\hat{y} = a + \frac{b}{x}$;
- полулогарфмическая функция: $\hat{y} = a + b \cdot ln\, x$;

2. Регрессии, *нелинейные по оцениваемым параметрам*, например

- степенная: $\hat{y} = a \cdot x^b$;
- показательная: $\hat{y} = a \cdot b^x$;

– экспоненциальная: $\hat{y} = e^{a+bx}$;

– обратная: $\hat{y} = \frac{1}{a+bx}$.

Регрессии этого класса в свою очередь делятся на два типа:

– нелинейные модели внутренне линейные (приводятся к линейному виду с помощью соответствующих преобразований, например, логарифмированием): показательная, степенная, обратная экспоненциальная и др.

– нелинейные модели внутренне нелинейные (к линейному виду не приводятся): $\hat{y} = a + b \cdot x^c$.

Оценка параметров нелинейной парной регрессии

Нелинейные модели регрессии сначала приводятся к линейному виду, а дальнейшая оценка параметров производится с помощью метода наименьших квадратов.

Регрессии 1-го класса (нелинейные по включенным переменным) приводятся к линейному виду простой заменой переменных.

Например, полулогарифмическая функция $\hat{y} = a + b \cdot ln\, x$ приводится к линейному виду с помощью замены: $z = ln\, x$. Получим $\hat{y}_z = a + b \cdot z$. Тогда формулы для определения параметров уравнения будут выглядеть следующим образом

$$b = \frac{cov(z,y)}{\sigma_z^2} = \frac{\overline{zy} - \overline{z} \cdot \overline{y}}{\overline{z^2} - \overline{z}^2},$$

$$a = \overline{y} - b\overline{z}.$$

Аналогичным образом приводятся к линейному виду зависимости $\hat{y} = a + \frac{b}{x}$, $\hat{y} = a + b \cdot \sqrt{x}$ и другие.

Среди нелинейных моделей достаточно часто используется степенная функция $y = a \cdot x^b \cdot \varepsilon$, которая приводится к линейному виду логарифмированием:

$$ln\, y = ln(a \cdot x^b \cdot \varepsilon),$$

$$ln\, y = ln\, a + b \cdot l\, n\, x + ln\, \varepsilon.$$

Введём обозначения $ln\, y = Y$; $ln\, a = A$; $ln\, x = X$; $ln\, \varepsilon = E$. Получим линейное уравнение $Y = A + bX + E$, оценить параметры которого можно методом наименьших квадратов:

$$b = \frac{cov(X,Y)}{\sigma_X^2} = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - \overline{X}^2},$$

$$A = \overline{Y} - b\overline{X}.$$

Чтобы получить исходное уравнение выполняют потенцирование.

Параметр b в степенной модели является коэффициентом эластичности.

Коэффициент эластичности

В экономических исследованиях широкое применение находит такой показатель как *коэффициент эластичности*, вычисляемый по формуле

$$Э = y' \frac{x}{y}.$$

Коэффициент эластичности показывает на сколько процентов изменится результат y при изменении фактора x на 1% от своего номинального значения.

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

$$\overline{Э} = y' \frac{\bar{x}}{\bar{y}}.$$

Оценка тесноты связи

Уравнение нелинейной регрессии, так же, как и в случае линейной зависимости, дополняется показателем тесноты связи. В данном случае это индекс корреляции:

$$\rho_{xy} = \sqrt{1 - \frac{\sigma_{ост}^2}{\sigma_y^2}},$$

где $\sigma_y^2 = \frac{1}{n}\sum_i (y_i - \bar{y})^2$ – общая дисперсия результативного признака y,

$\sigma_{ост}^2 = \frac{1}{n}\sum_i (y_i - \hat{y}_i)^2$ – остаточная дисперсия.

Величина данного показателя находится в пределах: $0 \le \rho_{xy} \le 1$. Чем ближе значение индекса корреляции к единице, тем теснее связь рассматриваемых признаков, тем более надежно уравнение регрессии.

Оценка качества построенной модели

Квадрат индекса корреляции носит название индекса детерминации и имеет тот же смысл, что и в линейной регрессии, т.е. характеризует долю дисперсии результативного признака y, объясняемую регрессией, в общей дисперсии результативного признака:

$$\rho_{xy}^2 = 1 - \frac{\sigma_{ост}^2}{\sigma_y^2} = \frac{\sigma_{факт}^2}{\sigma_y^2},$$

$\sigma_{факт}^2 = \frac{1}{n}\sum_i (\hat{y}_i - \bar{y})^2$ – факторная, объяснённая дисперсия.

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

Как и в линейном случае о качестве нелинейного уравнения регрессии можно также судить и по средней ошибке аппроксимации, которая, вычисляется по той же формуле:

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

Оценка значимости уравнения нелинейной регрессии

Оценка значимости уравнения регрессии в целом производится на основе *F*-критерия Фишера:

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

n – число наблюдений, *m* – число параметров при переменной *x*.

Фактическое значение *F*-критерия Фишера сравнивается с табличным значением $F_{\text{таб}}(\alpha; k_1; k_2)$ 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.

Если нелинейное относительно объясняющей переменной уравнение регрессии при линеаризации принимает форму линейного уравнения парной регрессии, то величина линейного коэффициента детерминации в этом случае совпадает с величиной индекса детерминации:

$$r_{zy}^2 = \rho_{xy}^2,$$

где *z* – преобразованная величина признака-фактора.

Следовательно, значения *F*-критерия Фишера для нелинейного уравнения регрессии и линеаризованного будут также совпадать.

Решение типовых задач

Задача 1. По данным проведенного опроса восьми групп семей известны данные связи расходов населения на продукты питания с уровнем доходов семьи (табл. 3).

Таблица 3

Расходы на продукты питания, *y*, тыс. руб.	0,9	1,2	1,8	2,2	2,6	2,9	3,3	3,8
Доходы семьи, *x*, тыс. руб.	1,2	3,1	5,3	7,4	9,6	11,8	14,5	18,7

1. Постройте поле корреляции и сформулируйте гипотезу о форме связи.

2. Постройте уравнение линейной регрессии.

3. Рассчитайте коэффициент линейной корреляции и коэффициент детерминации.

4. Оцените с помощью средней ошибки аппроксимации качество уравнения.

5. Дайте с помощью среднего коэффициента эластичности сравнительную оценку силы связи фактора с результатом.

6. С помощью *t*-критерия Стьюдента оцените статистическую значимость параметров регрессии (найдите их доверительные интервалы) и коэффициента корреляции.

7. Оцените с помощью *F*-критерия Фишера значимость уравнения.

8. Рассчитайте прогнозное значение результата, если прогнозное значение фактора увеличится на 10% от его среднего уровня. Оцените точность прогноза, рассчитав ошибку прогноза и его доверительный интервал.

9. Оцените полученные результаты, оформите выполненное задание в виде отчета.

Решение

1. Построим поле корреляции (рис. 1). По графику видно, что точки выстраиваются практически в прямую линию. Можно предположить, что *x* и *y* связаны линейной зависимостью.

Для удобства дальнейших вычислений составим расчётную таблицу (табл. 4). Столбцы 2 и 3 заполняются исходными данными. Как вычисляются столбцы 4 – 6 понятно из их заголовков.

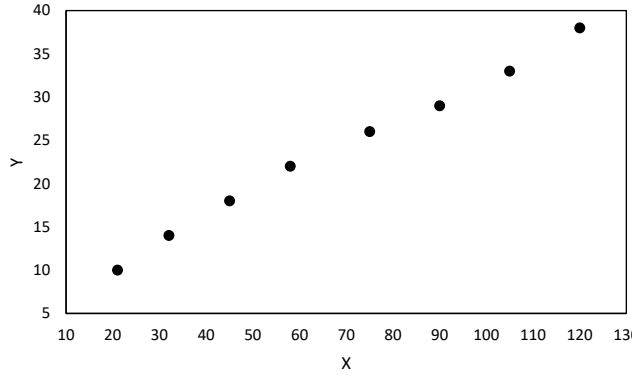


Рис. 1. Поле корреляции.

Средние значения рассчитываются по следующим формулам:

$$\overline{x} = \frac{\sum x_i}{n}, \qquad \overline{y} = \frac{\sum y_i}{n}, \qquad \overline{x^2} = \frac{\sum x_i^2}{n}, \qquad \overline{xy} = \frac{\sum x_i y_i}{n}.$$

Таблица 4

№	x	y	$x \cdot y$	x^2	y^2	$\hat{y}_x$	$y - \hat{y}$	$(y - \hat{y})^2$	A_i, %
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	21	10	210	441	100	10,983	-0,983	0,9655	9,83
2	32	14	448	1024	196	13,955	0,045	0,0020	0,32
3	45	18	810	2025	324	17,468	0,532	0,2834	2,96
4	58	22	1276	3364	484	20,980	1,020	1,0397	4,63
5	75	26	1950	5625	676	25,574	0,426	0,1815	1,64
6	90	29	2610	8100	841	29,627	-0,627	0,3932	2,16
7	105	33	3465	11025	1089	33,680	-0,680	0,4627	2,06
8	120	38	4560	14400	1444	37,733	0,267	0,0711	0,70
Итого	546	18,7	15329	46004	5154	190,000	0,000	3,3992	24,30
Среднее значение	68,25	23,75	1916,13	5750,50	644,25	23,750	-	0,4249	3,04

$\sigma_x^2 = \overline{x^2} - \overline{x}^2 = 5750{,}5 - 68{,}25^2 = 1092{,}44$ – дисперсия признака x,

$\sigma_x = 33{,}05$ – среднеквадратическое отклонение x.

$\sigma_y^2 = \overline{y^2} - \overline{y}^2 = 644{,}25 - 23{,}75^2 = 80{,}19$ – дисперсия признака y,

$\sigma_y = 8{,}95$ – среднеквадратическое отклонение y.

2. Рассчитаем параметры линейного уравнения парной регрессии $\hat{y} = a + b \cdot x$:

$$b = \frac{\overline{xy} - \overline{xy}}{\overline{x^2} - \overline{x}^2} = \frac{1916{,}13 - 68{,}25 \cdot 23{,}75}{1092{,}44} = 0{,}270;$$

$$a = \overline{y} - b\overline{x} = 2{,}34 - 0{,}169 \cdot 8{,}95 = 5{,}308.$$

Получили уравнение: $\hat{y} = 5{,}308 + 0{,}270 \cdot x$.

Параметр $b = 0{,}270$ означает, что с увеличением дохода семьи на 1 тыс. руб. расходы на продукты питания увеличиваются в среднем на 0,270 тыс. руб. (или 270 руб.)

3. Коэффициент линейной корреляции

$$r_{xy} = b \cdot \frac{\sigma_x}{\sigma_y} = 0{,}270 \cdot \frac{33{,}05}{8{,}95} = 0{,}997.$$

Близость коэффициента корреляции к 1 указывает на очень тесную прямую линейную связь между признаками.

Коэффициент детерминации $r_{xy}^2 = 0{,}995$ показывает, что уравнением регрессии объясняется 99,5% дисперсии результативного признака, а на долю прочих факторов приходится лишь 0,5%.

Или

Вариация результата (расходов на продукты) на 99,5% объясняется вариацией фактора (доходов семьи).

Далее рассчитываем столбцы 7-9.

Столбец 7 – теоретические значения $\hat{y}$ рассчитываются путём подстановки значений x в полученное уравнение:

например,

$\hat{y}_1 = 5{,}308 + 0{,}270 \cdot 21 = 10{,}983$ и т.д.

Столбцы 8, 9 – отклонение фактических значений от теоретических и квадрат отклонения соответственно:

$y_1 - \hat{y}_1 = 10 - 10{,}983 = -0{,}983$ и т.д.

$(y_1 - \hat{y}_1)^2 = (-0{,}983)^2 = 0{,}9655$ и т.д.

Столбец 10:

$$A_i = \left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot 100\%.$$

$A_1 = \left|\frac{y_1 - \hat{y}_1}{y_1}\right| \cdot 100\% = \left|\frac{-0{,}983}{10}\right| \cdot 100\% = 9{,}83\%$ (значения столбца 8 делятся на значения столбца 3). Так как в формуле модуль, то *все значения должны быть положительными*.

4. Средняя ошибка аппроксимации $\bar{A} = \frac{1}{n}\sum_i \left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot 100\%$. Её значение смотрим в таблице – среднее значение столбца 10.

$\bar{A} = 3{,}04\% < 8 – 10\%$ – качество модели очень хорошее.

5. Средний коэффициент эластичности равен

$$\overline{Э} = y' \frac{\bar{x}}{\bar{y}} = 0{,}270 \frac{68{,}25}{23{,}75} = 0{,}78.$$

Следовательно, при увеличении доходов семьи на 1% от своего среднего значения расходы на продукты питания в среднем по совокупности увеличатся на $0{,}78\%$.

6. Для оценки статистической значимости параметров регрессии и коэффициента корреляции рассчитаем *t*-критерий Стьюдента и доверительные интервалы каждого из показателей. Рассчитаем случайные ошибки параметров линейной регрессии и коэффициента корреляции. Для расчёта $S^2_{ост}$ сумма столбца 9 делится на *n* – 2.

$$S^2_{ост} = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-2} = \frac{3{,}3992}{8-2} = 0{,}567.$$

$$m_b = \frac{S_{ост}}{\sigma_x \cdot \sqrt{n}} = \frac{\sqrt{0{,}567}}{33{,}05 \cdot \sqrt{8}} = 0{,}0081.$$

$$m_a = S_{ост} \frac{\sqrt{\sum x^2}}{\sigma_x \cdot n} = \frac{\sqrt{0{,}567 \cdot 46004}}{33{,}05 \cdot 8} = 0{,}6105.$$

$$m_r = \sqrt{\frac{1 - r^2_{xy}}{n-2}} = \sqrt{\frac{1 - 0{,}982}{6}} = 0{,}0297.$$

Фактические значения *t* -статистик:

$$t_b = \frac{b}{m_b} = \frac{0{,}270}{0{,}0081} = 33{,}561; \quad t_a = \frac{a}{m_a} = \frac{5{,}308}{0{,}6105} = 8{,}694;$$

$$t_r = \frac{r}{m_r} = \frac{0{,}991}{0{,}0541} = 33{,}561.$$

Табличное значение *t*-критерия Стьюдента при $\alpha = 0{,}05$ и числе степеней свободы $\nu = n - 2 = 6$ есть $t_{таб} = 2{,}4469$. Так как

$t_b > t_{таб}$, то параметр *b* (коэффициент регрессии) статистически значим,

$t_a > t_{таб}$ – параметр *a* статистически значим,

и $t_r > t_{таб}$ – коэффициент корреляции статистически значим.

Рассчитаем доверительные интервалы для параметров регрессии:

$$b - t_{таб} \cdot m_b \le b^* \le b + t_{таб} \cdot m_b,$$

$$0{,}270 - 2{,}4469 \cdot 0{,}0081 \le b^* \le 0{,}270 + 2{,}4469 \cdot 0{,}0081,$$

$$0{,}251 \le b^* \le 0{,}290.$$

$$a - t_{таб} \cdot m_a \le a^* \le a + t_{таб} \cdot m_a$$
$$5{,}308 - 2{,}4469 \cdot 0{,}6105 \le a^* \le 5{,}308 + 2{,}4469 \cdot 0{,}6105,$$
$$3{,}814 \le a^* \le 6{,}802.$$

7. Оценим качество уравнения регрессии в целом с помощью *F*-критерия Фишера. Сосчитаем фактическое значение *F*-критерия:

$$F = \frac{r_{xy}^2}{1 - r_{xy}^2} \cdot (n-2) = \frac{0{,}995}{1 - 0{,}995} \cdot 6 = 1126{,}34.$$

Табличное значение при $\alpha = 0{,}05$, $k_1 = 1$ и $k_2 = n - 2 = 6$: $F_{\text{таб}}(0{,}05, 1, 6) = 5{,}99$. Так как $F_{\text{факт}} > F_{\text{таб}}$, то признается статистическая значимость уравнения в целом.

8. Найдем прогнозное значение результативного фактора при значении признака-фактора, составляющем 110% от среднего уровня:

$$x_p = 1{,}1 \cdot \bar{x} = 1{,}1 \cdot 68{,}25 = 75{,}075,$$

т.е. найдем расходы на питание, если доходы семьи составят 75,075 тыс. руб.

$$y_p = 5{,}308 + 0{,}270 \cdot 75{,}075 = 25{,}594 \text{ (тыс. руб.)}$$

Значит, если доходы семьи составят 75,075 тыс. руб., то расходы на продукты питания будут равны 25,594 тыс. руб.

Найдем доверительный интервал прогноза. Ошибка прогноза равна

$$m_{y_p} = S_{ост} \cdot \sqrt{1 + \frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{n \cdot \sigma_x^2}} =$$

$$= \sqrt{0{,}567 \cdot \left(1 + \frac{1}{8} + \frac{(75{,}075 - 68{,}25)^2}{8 \cdot 1092{,}44}\right)} = 0{,}8.$$

$$y_p - t_{таб} \cdot m_{y_p} \le y_p^* \le y_p + t_{таб} \cdot m_{y_p},$$
$$25{,}594 - 2{,}4469 \cdot 0{,}8 \le y_p^* \le 25{,}594 + 2{,}4469 \cdot 0{,}8,$$
$$23{,}636 \le y_p^* \le 27{,}552.$$

То есть, с вероятностью 0,95 прогнозируемая величина расходов на продукты питания при доходах семьи равных 75,075 тыс. руб. будет принадлежать интервалу от 23,636 до 27,552 тыс. руб.

Задача 2. Для исходных данных из задачи 1

1. Постройте уравнение нелинейной регрессии (функцию выберете самостоятельно).

2. Рассчитайте индекс корреляции и индекс детерминации.

3. Оцените с помощью средней ошибки аппроксимации качество уравнения.

4. Дайте с помощью среднего коэффициента эластичности сравнительную оценку силы связи фактора с результатом.

5. Оцените с помощью *F*-критерия Фишера значимость уравнения.

6. Сделайте выводы.

Решение

Линейная зависимость предполагает, что с ростом доходов семьи абсолютный прирост расходов на продукты питания будет постоянным. То есть при очень больших доходах расходы на продукты будут тоже очень большими. Более соответствует реальности рост расходов на продукты питания с замедляющимся темпом. Предположим, что между признаками нелинейная зависимость. Пусть она имеет следующий вид $\hat{y}_x = a + b \cdot \sqrt{x}$.

Приведём уравнение к линейному виду с помощью замены переменной $z = \sqrt{x}$ и найдём параметры линейного уравнения $\hat{y}_z = a + b \cdot z$ методом наименьших квадратов. Для этого составим аналогичную расчётную таблицу (табл. 5), но с новой переменной *z*.

Таблица 5

	x	z	y	$z \cdot y$	z^2	y^2	$\hat{y}_z$	$y - \hat{y}_z$	$(y - \hat{y}_z)^2$	A_i, %
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
1	21	4,58	10	45,83	21	100	9,309	0,691	0,4777	6,91
2	32	5,66	14	79,20	32	196	13,863	0,137	0,0187	0,98
3	45	6,71	18	120,75	45	324	18,320	-0,320	0,1025	1,78
4	58	7,62	22	167,55	58	484	22,168	-0,168	0,0282	0,76
5	75	8,66	26	225,17	75	676	26,596	-0,596	0,3550	2,29
6	90	9,49	29	275,12	90	841	30,100	-1,100	1,2100	3,79
7	105	10,25	33	338,15	105	1089	33,322	-0,322	0,1040	0,98
8	120	10,95	38	416,27	120	1444	36,322	1,678	2,8162	4,42
Итого	546	63,91	190	1668,02	546	5154	190,00	0,000	5,11	21,91
Среднее значение	68,25	7,99	23,75	208,50	68,25	644,25	23,75	-	0,6390	2,74

$\sigma_z^2 = \overline{z^2} - \overline{z}^2 = 68{,}25 - 7{,}99^2 = 4{,}43$ – дисперсия *z*.

$\sigma_y^2 = 80{,}19$ – дисперсия *y*.

$\sigma_{ост}^2 = \frac{1}{n}\sum_i (y_i - \hat{y}_i)^2 = 0{,}6390$ – остаточная дисперсия (среднее значение столбца 10).

1. Найдём параметры уравнения:

$$b = \frac{\overline{zy} - \overline{z} \cdot \overline{y}}{\sigma_z^2} = \frac{208{,}50 - 7{,}99 \cdot 23{,}75}{4{,}43} = 4{,}239,$$

$$a = \overline{y} - b \cdot \overline{z} = 23{,}75 - 4{,}24 \cdot 7{,}99 = -10{,}119.$$

Получим уравнение регрессии

$\hat{y}_z = -10{,}119 + 4{,}239 \cdot z$ или в исходном виде $\hat{y}_x = -10{,}119 + 4{,}239 \cdot \sqrt{x}$.

В данном нелинейном уравнении параметры не интерпретируются.

Теперь заполняем столбцы 8–11 таблицы.

Столбец 8 – теоретические значения $\hat{y}_z$ рассчитываются путём подстановки значений z в полученное уравнение:

например,

$\hat{y}_1 = -10{,}119 + 4{,}239 \cdot 21 = 9{,}309$ и т.д.

Столбцы 9, 10 – отклонение фактических значений от теоретических и квадрат отклонения соответственно:

$y_1 - \hat{y}_1 = 10 - 9{,}309 = 0{,}691$ и т.д.

$(y_1 - \hat{y}_1)^2 = (0{,}691)^2 = 0{,}4777$ и т.д.

2. Индекс корреляции

$$\rho_{xy} = \sqrt{1 - \frac{\sigma_{ост}^2}{\sigma_y^2}} = \sqrt{1 - \frac{0{,}6390}{80{,}19}} = 0{,}996.$$

Связь между показателями очень тесная.

Индекс детерминации $\rho_{xy}^2 = 0{,}992$ показывает, что 99,2% дисперсии результативного признака (расходы на продукты питания) объясняется вариацией признак-фактора (доходы семьи). А на долю прочих факторов приходится 0,08%.

3. Средняя ошибка аппроксимации $\bar{A} = 2{,}74\%$ (среднее значение столбца 11) существенно меньше 8%, следовательно качество подбора кривой очень хорошее.

4. Средний коэффициент эластичности:

$$y' = (a + b \cdot \sqrt{x})' = \frac{b}{2\sqrt{x}}.$$

$$\overline{Э} = y' \frac{\bar{x}}{\bar{y}} = \frac{b}{2\sqrt{\bar{x}}} \cdot \frac{\bar{x}}{\bar{y}} = \frac{b \cdot \sqrt{\bar{x}}}{2 \cdot \bar{y}} = \frac{4{,}239 \cdot \sqrt{68{,}25}}{2 \cdot 23{,}75} = 0{,}74.$$

Следовательно, при увеличении доходов семьи на 1% от своего среднего значения расходы на продукты питания в среднем по совокупности увеличатся на 0,74%.

5. Критерий Фишера $F = \frac{0{,}992}{1-0{,}992} \cdot \frac{8-1-1}{1} = 622{,}4$. Фактическое значение больше табличного $F_{\text{таб}}(0{,}05; 1; 6) = 5{,}99$, значит уравнение статистически значимо.

Задача 3. Имеется следующая модель регрессии, характеризующая зависимость *y* от *x*: $y = 7 - 5 \cdot x + \varepsilon$.

Известно также, что $r = -0{,}6;\ n = 15$.

Задание:

1. Построить доверительный интервал для коэффициента регрессии:

а) с вероятностью 99%;

б) с вероятностью 95%.

2. Проанализировать результаты и пояснить причины их различий.

Решение

1. Доверительный интервал найдём по формуле

$$b - m_b \cdot t_{\text{таб}} \le b^* \le b + m_b \cdot t_{\text{таб}};$$

Стандартную ошибку m_b можно выразить из формулы для *t*-критерия: $m_b = \frac{b}{t_b}$; *t*-критерий выразить через *F*-критерий: $t_b^2 = F \to |t_b| = \sqrt{F}$.

А *F*-критерий, в свою очередь, найдём через квадрат коэффициента корреляции: $F = \frac{r^2}{1-r^2}(n-2) = \frac{(-0{,}6)^2}{1-(-0{,}6)^2}(15-2) = \frac{0{,}36}{1-0{,}36}13 = 7{,}3125$.

$t_b = \sqrt{7{,}3125} = 2{,}704$; $m_b = \frac{5}{2{,}704} = 1{,}849$ (так как стандартная ошибка не может быть отрицательной, то *b* берём без минуса).

Теперь можно вычислять доверительные интервалы.

а) вероятность 99% соответствует уровню значимости $\alpha = 0{,}01$: $t_{таб}(0{,}01; 13) = 3{,}0123$.

$$-5 - 1{,}849 \cdot 3{,}0123 \le b^* \le -5 + 1{,}849 \cdot 3{,}0123;$$

$$-10{,}570 \le b^* \le 0{,}570.$$

Так как концы доверительного интервала имеют значения разных знаков, т.е. ноль попадает в интервал, то при уровне значимости 0,01 коэффициент *b* равен нулю – статистически не значим.

б) вероятность 95% соответствует уровню значимости $\alpha = 0{,}05$: $t_{таб}(0{,}05; 13) = 2{,}1604$;

$$-5 - 1{,}849 \cdot 2{,}1604 \le b^* \le -5 + 1{,}849 \cdot 2{,}1604;$$

$$-8{,}995 \le b^* \le -1{,}005.$$

Концы доверительного интервал имеют значения одного знака, следовательно, при уровне значимости 0,05 коэффициент b статистически значим.

2. Различия полученных результатов объясняются величиной заданной вероятности – чем выше доверительная вероятность, тем шире границы доверительного интервала. Для одного уровня значимости параметр может быть значимым, а для другого тот же самый параметр оказывается не значимым.

Задача 4.

Зависимость уровня жизни от размера заработной платы характеризуется моделью:

$$\hat{y} = a + b \cdot x + c \cdot x^2$$

Результаты её использования представлены в таблице 6.

Таблица 6

Уровень жизни y, тыс. руб.	фактическая	12	8	13	15	16	11	12	9	11	9
	расчетная	10	10	13	14	15	12	13	10	10	9

Задание:

Оценить качество модели, определив среднюю ошибку аппроксимации, индекс корреляции и критерий Фишера.

Решение

Составим расчётную таблицу:

Таблица 7

№	Уровень жизни, тыс. руб., y		$y - \hat{y}$	$(y - \hat{y})^2$	A_i	y^2
	фактическая	расчетная				
1	12	10	2	4	0,1667	144
2	8	10	-2	4	0,25	64
3	13	13	0	0	0	169
4	15	14	1	1	0,0667	225
5	16	15	1	1	0,0625	256
6	11	12	-1	1	0,0909	121
7	12	13	-1	1	0,0833	144
8	9	10	-1	1	0,1111	81
9	11	10	1	1	0,0909	121
10	9	9	0	0	0	81
Сумма	116		0	14	0,9221	1406

Средняя ошибка аппроксимации

$\bar{A} = \frac{1}{10} 0{,}9221 * 100\% = 9{,}221\% < 10\%$ – качество модели удовлетворительное.

$$\sigma_y^2 = \overline{y^2} - \overline{y}^2 = 140{,}6 - 11{,}6^2 = 6{,}04;$$

$$\sigma_{ост}^2 = \frac{1}{n}\sum_i (y_i - \hat{y}_i)^2 = \frac{14}{10} = 1{,}4.$$

Индекс корреляции $\rho = \sqrt{1 - \frac{\sigma_{ост}^2}{\sigma_y^2}} = \sqrt{1 - \frac{1{,}4}{6{,}04}} = \sqrt{0{,}768} = 0{,}876$ – связь между показателями тесная;

$F = \frac{\rho^2}{1-\rho^2}\frac{n-m-1}{m} = \frac{0{,}768}{1-0{,}768}\frac{10-2-1}{2} = 11{,}6 > F_{\text{таб}}(0{,}05; 2; 10-2-1) = 4{,}74$ – уравнение статистически значимо.

Замечание. В данной модели два параметра при переменной *x* – это *b* и *c*, поэтому $m = 2$.

Построение парной регрессии с использованием табличного процессора MS Excel

Функция ЛИНЕЙН

Функция ЛИНЕЙН специально создана для оценки параметров линейной регрессии, а также для вывода регрессионной статистики (коэффициента детерминации, стандартных ошибок, F-критерия и др.)

Синтаксис функции

ЛИНЕЙН(известные_значения_y; [известные_значения_x]; [конст]; [статистика])

Известные_значения_y – диапазон, содержащий данные результативного признака (обязательный аргумент);

Известные_значения_x – диапазон, содержащий данные факторов независимого признака;

Конст – логическое значение, которое указывает на наличие или отсутствие свободного члена в уравнении (ИСТИНА – свободный член рассчитывается обычным образом; ЛОЖЬ – свободный член равен 0);

Статистика – логическое значение, которое указывает, выводить дополнительную информацию по регрессионному анализу или нет (ИСТИНА – дополнительная информация выводится, ЛОЖЬ – выводятся только оценки параметров уравнения).

В случае парной регрессии $\hat{y} = a + b \cdot x$ функция ЛИНЕЙН возвращает 10 статистик в следующем формате

Таблица 8

Значение коэффициента b	Значение параметра a
Стандартная ошибка b m_b	Стандартная ошибка a m_a
Коэффициент детерминации R^2	Оценка стандартного отклонения остатков $S_{\text{ост}}$
Значение F-статистики	Число степеней свободы
Регрессионная сумма квадратов $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	Остаточная сумма квадратов $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

Рассмотрим, как работает данная функция на примере данных из задачи 1.

Внесите исходные данные на лист MS Excel. Выделите диапазон размером 5×2 (5строк, 2 столбца). Введите формулу

«=ЛИНЕЙН(B2:B9;A2:A9;ИСТИНА;ИСТИНА)»

Или активируйте *Вставку функции* любым способом. В категории «Статистические» выберете функцию ЛИНЕЙН.

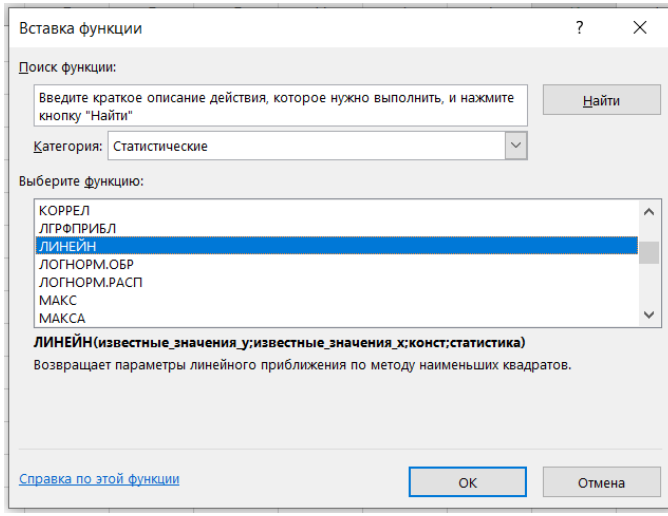


Рис. 2. Диалоговое окно Вставка функции.

Заполните аргументы функции:

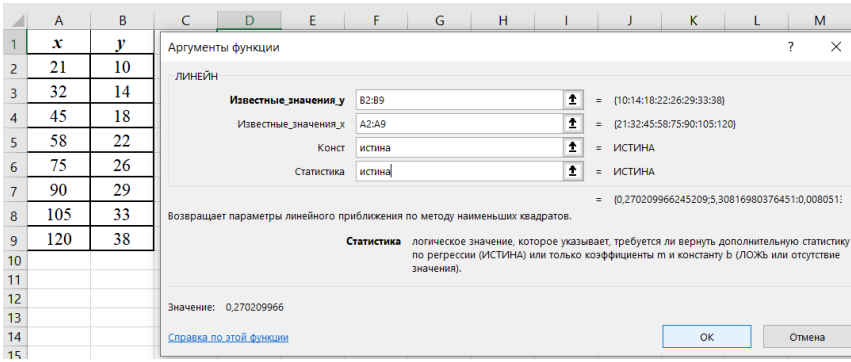


Рис. 3. Диалоговое окно функции ЛИНЕЙН.

Щёлкните по кнопке ОК. В левой верхней клетке выделенной области появится первый элемент итоговой таблицы. Чтобы получить все значения нажмите клавишу F2, а затем сочетание клавиш Ctrl+Shift+Enter. Получим

D	E
0,27021	5,30817
0,008051	0,610548
0,994701	0,75268
1126,337	6
638,1008	3,399165

Рис. 4. Результат функции ЛИНЕЙН.

Значение параметров a = 5,308 и b = 0,270; стандартные ошибки m_a = 0,6105, m_b = 0,0081; коэффициента детерминации r_{xy}^2 = 0,995; критерий Фишера F = 1126,34, $S_{ост}$ = 0,753.

Замечание. Можно сначала ввести формулу в ячейку, а затем выделить диапазон 5×2, нажать клавишу F2, сочетание клавиш Ctrl+Shift+Enter.

Инструмент Регрессия

Рассмотрим решение задач 1 и 2 с помощью инструмента *Регрессия* табличного процессора Excel.

Для построения уравнения линейной регрессии можно использовать надстройку **«Анализ данных»** табличного процессора Excel. Надстройка **«Анализ данных»** доступна на вкладке **Данные** в группе **Анализ**.

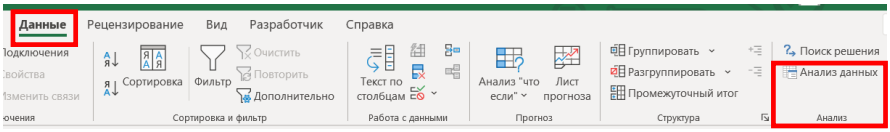


Рис. 5. Вкладка «Данные».

Если пакет анализа ещё не загружен, то его необходимо загрузить. В Excel 2010 и более поздних версиях выберите Файл → Параметры. В Excel 2007 нажмите кнопку Microsoft Office выберите «Параметры Excel» (рис. 6).

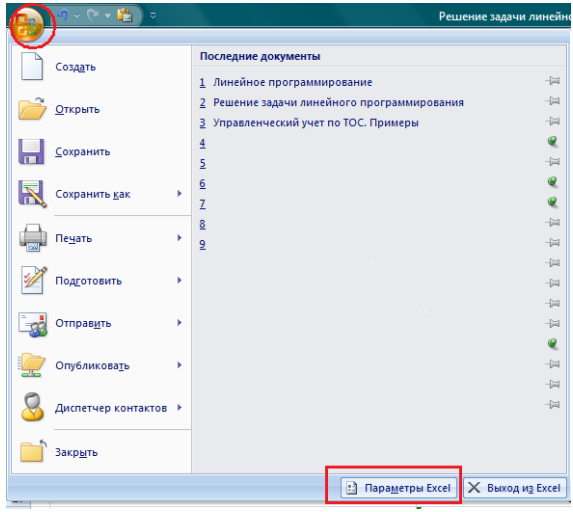


Рис. 6. Меню «Файл».

Выберите строку Надстройки, а затем в самом низу окна «Управление надстройками Microsoft Excel» выберите «Перейти» (рис. 7).

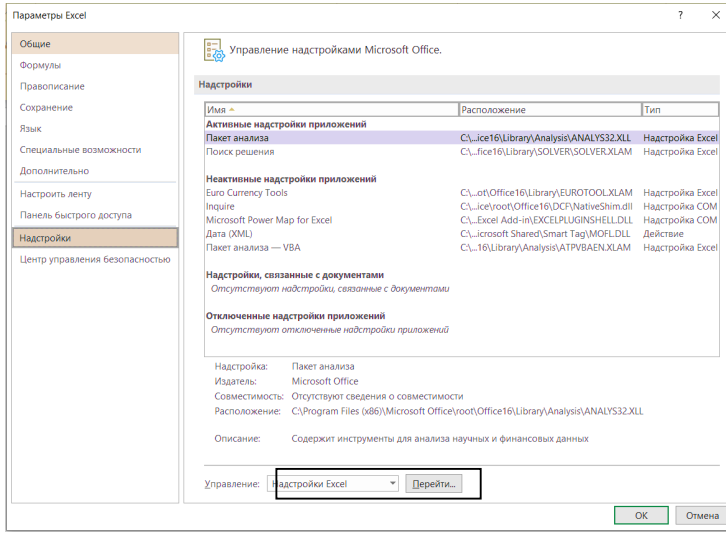


Рис. 7. Параметры Excel.

В окне «Надстройки» установите флажок «Поиск решения» и нажмите ОК (рис. 8).

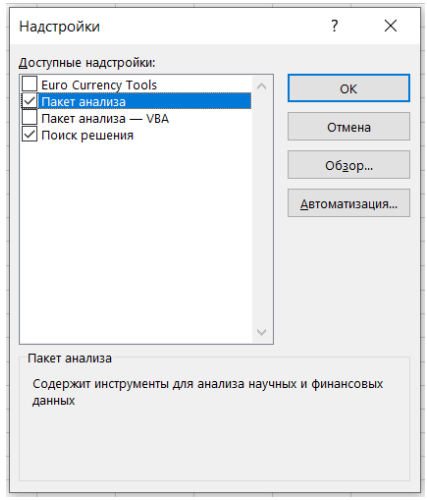


Рис. 8. Окно «Надстройки».

После вызова **Анализа данных** на экране появляется окно с перечнем инструментов анализа. Выбираем **Регрессия** и нажимаем кнопку **ОК** (рис. 9).

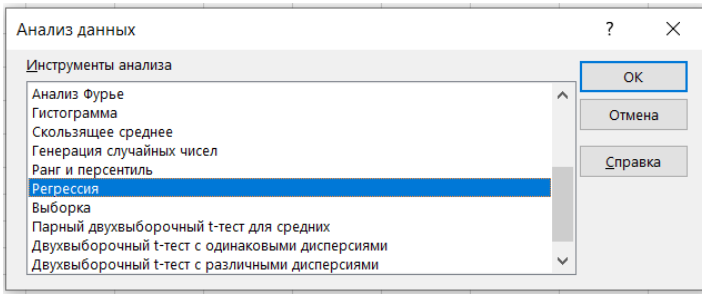


Рис. 9. Окно «Анализ данных».

На экране появится диалоговое окно, в котором задаются следующие параметры:

Входной интервал Y – диапазон, содержащий данные результативного признака;

Входной интервал X – диапазон, содержащий данные признака-фактора;

Метки – флажок, который указывает, содержит ли первая строка названия столбцов;

Константа – ноль – флажок, указывающий на наличие или отсутствие свободного члена в уравнении;

Уровень надежности – по умолчанию равен 95% и соответствует уровню значимости α = 0,05. Если вы хотите изменить это значение, установите флажок в соответствующем поле и измените значение.

В разделе «Параметры вывода» выберите, куда вы хотите вывести результаты регрессии:

Выходной интервал – достаточно указать левую верхнюю ячейку будущего диапазона;

Новый рабочий лист – можно указать произвольное имя нового листа (или не указывать, тогда результаты выводятся на вновь созданный лист).

Остатки – при включении вычисляется столбец, содержащий остатки для всех точек наблюдений (исходных данных) $\varepsilon_i = y_i - \hat{y}_i$, $i = 1, \ldots, n$.

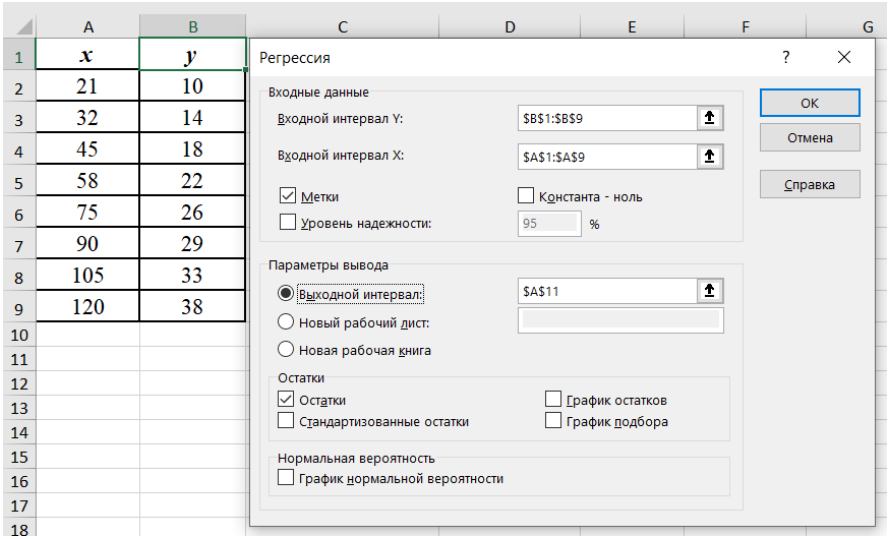
Стандартизованные остатки – при включении вычисляется столбец, содержащий стандартизованные остатки.

График остатков – при включении выводятся точечные графики остатков $\varepsilon_i = y_i - \hat{y}_i$, $i = 1, \ldots, n$ в зависимости от значений переменных x_j, $j = 1, \ldots, p$. Количество графиков равно числу переменных.

График подбора – при включении выводятся точечные графики предсказанных по построенной регрессии значений $\hat{y}_i$ от значений переменных x_j, при $j = 1, \ldots, p$.

Начнём с линейной модели. Для проведения расчётов предварительно необходимо внести исходные данные на лист Excel (2 столбца: *x*, *y*). Запускаем инструмент **Регрессия**. В появившемся окне зададим соответствующие диапазоны данных (рис. 10).

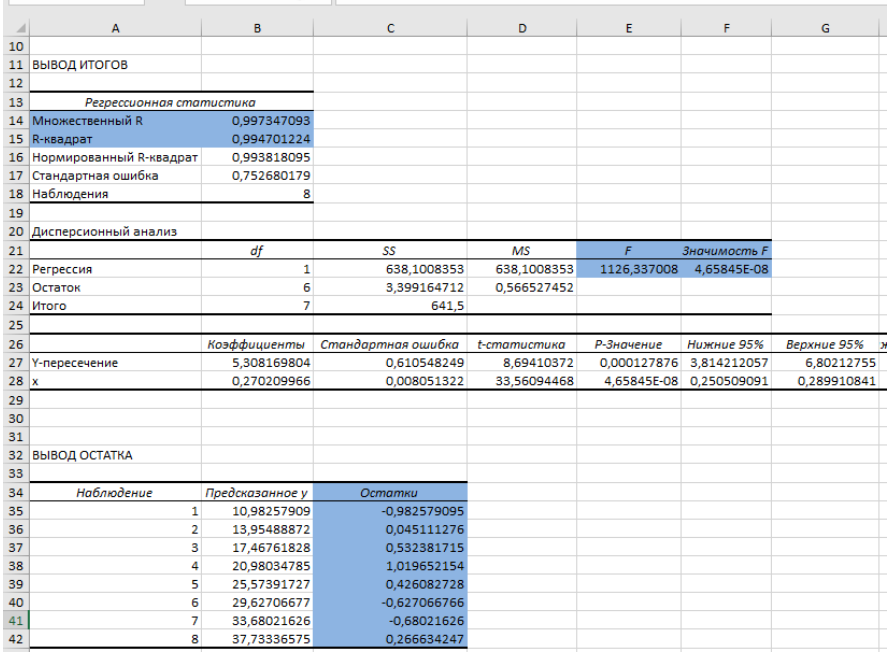
Столбцы выделяем вместе с заголовками и ставим флажок в поле «Метки». Итоги выведем на этот же рабочий лист, указав начальную ячейку, например A11. Установим флажок в поле «Остатки». Нажимаем **ОК**.



	A	B
1	*x*	*y*
2	21	10
3	32	14
4	45	18
5	58	22
6	75	26
7	90	29
8	105	33
9	120	38

Рис. 10. Диалоговое окно инструмента Регрессия для линейной модели.

Получим следующие результаты.



	A	B	C	D	E	F	G
10							
11	ВЫВОД ИТОГОВ						
12							
13	*Регрессионная статистика*						
14	Множественный R	0,997347093					
15	R-квадрат	0,994701224					
16	Нормированный R-квадрат	0,993818095					
17	Стандартная ошибка	0,752680179					
18	Наблюдения	8					
19							
20	Дисперсионный анализ						
21		*df*	*SS*	*MS*	*F*	*Значимость F*	
22	Регрессия	1	638,1008353	638,1008353	1126,337008	4,65845E-08	
23	Остаток	6	3,399164712	0,566527452			
24	Итого	7	641,5				
25							
26		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
27	Y-пересечение	5,308169804	0,610548249	8,69410372	0,000127876	3,814212057	6,80212755
28	x	0,270209966	0,008051322	33,56094468	4,65845E-08	0,250509091	0,289910841
29							
30							
31							
32	ВЫВОД ОСТАТКА						
33							
34	*Наблюдение*	*Предсказанное y*	*Остатки*				
35	1	10,98257909	-0,982579095				
36	2	13,95488872	0,045111276				
37	3	17,46761828	0,532381715				
38	4	20,98034785	1,019652154				
39	5	25,57391727	0,426082728				
40	6	29,62706677	-0,627066766				
41	7	33,68021626	-0,68021626				
42	8	37,73336575	0,266634247				

Рис. 11. Вывод итогов инструмента Регрессия для линейной модели.

Из первой таблицы получим коэффициент корреляции $r_{xy} = 0{,}997$ и коэффициента детерминации $r_{xy}^2 = 0{,}995$.

Из второй таблицы (дисперсионный анализ) получим фактическое значение критерия Фишера: $F = 1126{,}34$.

Табличное значение можно получить с помощью встроенной функции MS Excel F.ОБР.ПХ или для более ранних версий Excel функцию FРАСПОБР. В свободной ячейке набираем формулу «=F.ОБР.ПХ(0,05;1;6)» либо активируем *Вставку функции* и в категории «Статистические» выбираем функцию F.ОБР.ПХ и заполняем аргументы функции в диалоговом окне (рис. 12): **вероятность** = 0,05, **степени_свободы1** = 1, **степени_свободы2** = 6. В результате получаем $F_{\text{таб}} = 5{,}9874$.

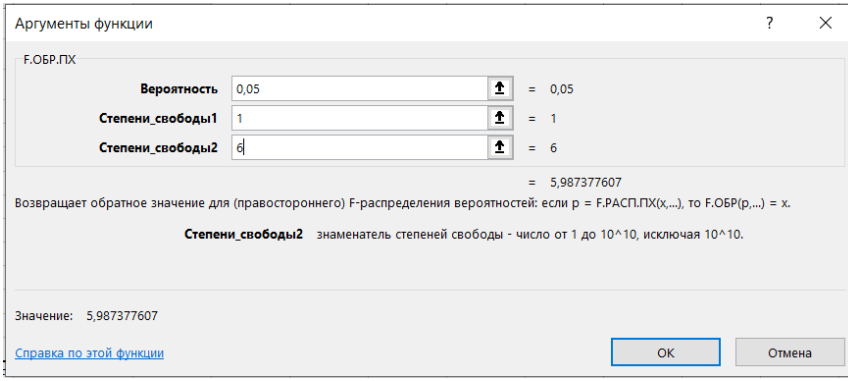


Рис. 12. Диалоговое окно функции F.ОБР.ПХ.

Число в столбце «*Значимость F*» – уровень значимости, соответствующий вычисленной величине F-критерия и равный вероятности того, что расчетное значение F-критерия меньше или равно табличному. Если вероятность меньше уровня значимости α, то построенная регрессия является значимой.

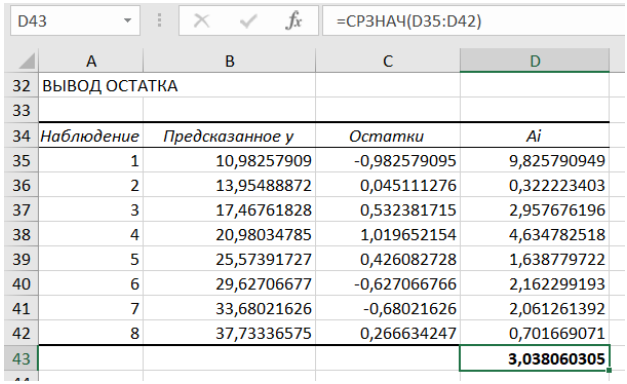
В нашем случае число 4,65845E-08 = $4{,}65845 \cdot 10^{-8}$ = 0,0000000465845 меньше, чем 0,05 из чего следует статистическая значимость уравнения в целом. Т.е. можно обойтись и без табличного значения F-критерия.

Из третьей таблицы получим значение параметров $a = 5{,}308$ (строка Y-пересечение) и $b = 0{,}270$ (строка x); стандартные ошибки $m_a = 0{,}6105$, $m_b = 0{,}0081$, t-статистики $t_a = 8{,}694$, $t_b = 33{,}561$ и доверительный интервал (*Нижние 95%*, *Верхние 95%*).

Параметр *P-Значение* используется для оценки статистической значимости параметров регрессии. Если P-значение меньше заданного уровня значимости α, то параметр является значимым на этом уровне значимости. В противном случае незначим на данном уровне значимости α. В нашем случае числа в столбце P-Значение меньше 0,05, следовательно, параметры a и b статистически значимы. Это удобно, так как не требуется смотреть табличное значение.

Найдены все параметры и характеристики уравнения регрессии за исключением средней ошибки аппроксимации.

В четвёртой таблице в столбце *Остатки* рассчитаны значения $y_i - \hat{y}_i$. Воспользуемся этими данными для расчёта средней ошибки аппроксимации. Для этого необходимо каждое значение i-го остатка разделить на y_i, взять по модулю, воспользовавшись функцией «ABS», и умножить на 100. Добавим к четвертой таблице столбец A_i и выполним эти расчёты. В строке первого наблюдения внесём формулу «=ABS(C35/B2)*100». Скопируем для остальных наблюдений и посчитаем среднее значение полученного столбца с помощью функции «СРЗНАЧ».



	A	B	C	D
32	ВЫВОД ОСТАТКА			
33				
34	*Наблюдение*	*Предсказанное y*	*Остатки*	*Ai*
35	1	10,98257909	-0,982579095	9,825790949
36	2	13,95488872	0,045111276	0,322223403
37	3	17,46761828	0,532381715	2,957676196
38	4	20,98034785	1,019652154	4,634782518
39	5	25,57391727	0,426082728	1,638779722
40	6	29,62706677	-0,627066766	2,162299193
41	7	33,68021626	-0,68021626	2,061261392
42	8	37,73336575	0,266634247	0,701669071
43				**3,038060305**

Рис. 13. Вычисление средней ошибки аппроксимации.

Это и будет величина средней ошибки аппроксимации: $\bar{A} = 3{,}04\%$.

Все результаты совпадают с полученными выше вручную.

Теперь перейдём к модели $\hat{y}_x = a + b \cdot \sqrt{x}$. Скопируем исходные данные на новый лист Excel и добавим столбец $z = \sqrt{x}$. Значения новой переменной z можно получить, воспользовавшись математической функцией КОРЕНЬ, которая возвращает значение квадратного корня.

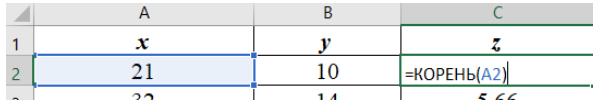


Рис. 14. Математическая функция КОРЕНЬ.

Или использовать оператор возведения в степень «^».

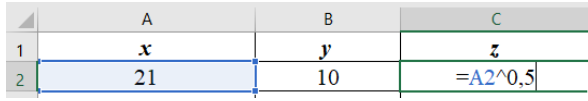


Рис. 15. Вычисление квадратного корня с помощью оператора возведения в степень.

Теперь запускаем инструмент **Регрессия**. В диалоговом окне задаём соответствующие диапазоны данных (рис. 16). В окне редактирования *Входной интервал Y*, как и в предыдущем случае, указываем диапазон столбца *y*. А в окне редактирования *Входной интервал X* – диапазон новой переменной *z*. Результаты выведем на текущий лист. Установим флажок в поле «Остатки». Нажимаем **ОК**.

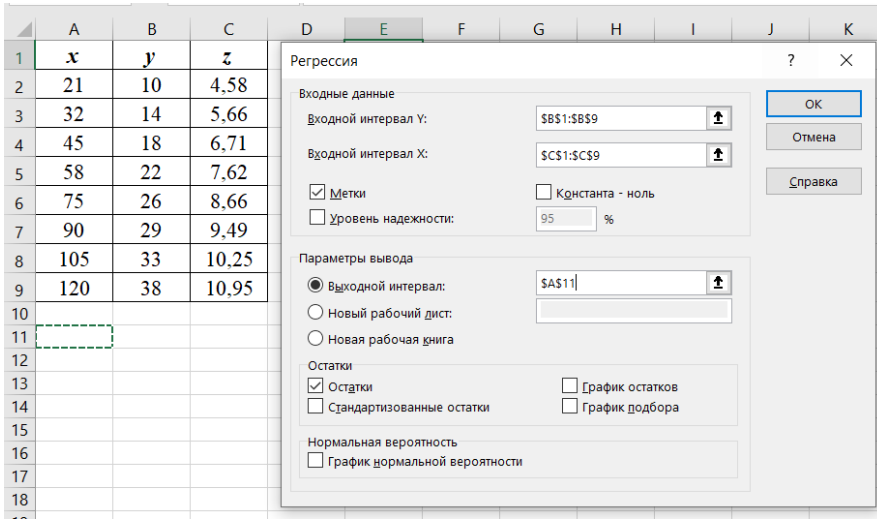


Рис. 16. Диалоговое окно инструмента Регрессия для нелинейной модели.

Получим следующие результаты, представленные на рисунке 17.

	A	B	C	D	E	F
10						
11	ВЫВОД ИТОГОВ					
12						
13	*Регрессионная статистика*					
14	Множественный R	0,996007361				
15	R-квадрат	0,992030664				
16	Нормированный R-квадрат	0,990702441				
17	Стандартная ошибка	0,923068193				
18	Наблюдения	8				
19						
20	Дисперсионный анализ					
21		*df*	*SS*	*MS*	*F*	*Значимость F*
22	Регрессия	1	636,3876707	636,3876707	746,885769	1,58642E-07
23	Остаток	6	5,112329331	0,852054889		
24	Итого	7	641,5			
25						
26		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*
27	Y-пересечение	-10,11863017	1,281534061	-7,895716919	0,000218815	-13,25443105
28	z	4,239414861	0,15512394	27,32921091	1,58642E-07	3,859840255
29						
30						
31						
32	ВЫВОД ОСТАТКА					
33						
34	*Наблюдение*	*Предсказанное y*	*Остатки*			
35	1	9,308809336	0,691190664			
36	2	13,8631218	0,136878196			
37	3	18,32022927	-0,320229274			
38	4	22,16779151	-0,167791515			
39	5	26,5957795	-0,5957795			
40	6	30,09999055	-1,099990553			
41	7	33,32244519	-0,322445188			
42	8	36,32183283	1,678167169			

Рис. 17. Вывод итогов инструмента Регрессия для нелинейной модели.

Из первой таблицы получим коэффициент корреляции $r_{zy} = 0{,}997$ (индекс корреляции $\rho_{xy} = 0{,}997$) и коэффициента детерминации $r_{zy}^2 = 0{,}995$ (индекс детерминации $\rho_{xy}^2 = 0{,}995$).

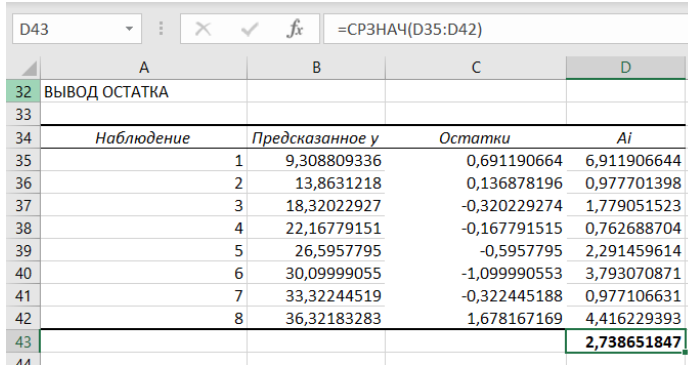
Из второй таблицы (дисперсионный анализ) получим фактическое значение критерия Фишера: $F = 746{,}88$.

Из третьей таблицы получим значение параметров $a = -10{,}119$ (строка Y-пересечение) и $b = 4{,}239$ (строка x); стандартные ошибки $m_a = 1{,}2815$, $m_b = 0{,}1551$, t-статистики $t_a = 7{,}8957$, $t_b = 27{,}3292$ и доверительный интервал (*Нижние 95%*, *Верхние 95%*).

Найдём среднюю ошибку аппроксимации.

Добавим к четвертой таблице столбец A_i и в строке первого наблюдения внесём формулу «=ABS(C35/B2)*100». Скопируем формулу для остальных наблюдений и посчитаем среднее значение полученного столбца с помощью функции «СРЗНАЧ» (рис. 18).

Средняя ошибка аппроксимации: $\bar{A} = 2{,}74\%$.



D43 | =СРЗНАЧ(D35:D42)

	A	B	C	D
32	ВЫВОД ОСТАТКА			
33				
34	*Наблюдение*	*Предсказанное y*	*Остатки*	*Ai*
35	1	9,308809336	0,691190664	6,911906644
36	2	13,8631218	0,136878196	0,977701398
37	3	18,32022927	-0,320229274	1,779051523
38	4	22,16779151	-0,167791515	0,762688704
39	5	26,5957795	-0,5957795	2,291459614
40	6	30,09999055	-1,099990553	3,793070871
41	7	33,32244519	-0,322445188	0,977106631
42	8	36,32183283	1,678167169	4,416229393
43				**2,738651847**

Рис. 18. Вычисление средней ошибки аппроксимации для нелинейной модели.

Все результаты совпадают с полученными выше при ручном счёте.

Задачи для самостоятельного решения

1. Получена зависимость между затратами y (тыс. руб.) и объемом выпуска продукции x (ед.)

$$\hat{y} = 35000 + 0{,}58 \cdot x.$$

Что показывает величина коэффициента регрессии?

2. Зависимость расходов населения на покупку товаров и оплату услуг (y, в процентах от общего объема денежных доходов) от среднедушевых денежных доходов населения (x, руб./мес.) описывается соотношением

$$\hat{y} = 105{,}88 - 0{,}0007 \cdot x.$$

Что показывает величина коэффициента регрессии?

3. По 10 заводам региона, выпускающих один и тот же вид продукции, исследуется зависимость себестоимости единицы продукции y (тыс. руб.) от уровня технической оснащенности x (тыс. руб.). Доля остаточной дисперсии в общей составила 17%. Оцените значимость уравнения линейной регрессии.

4. По совокупности 12 предприятий, производящих однородную продукцию, изучается зависимость между выпуском продукции x (тыс. ед.) и затратами на производство y (млн. руб.) При оценке модели линейной регрессии были получены следующие промежуточные результаты: $\sum_i (y_i - \bar{y})^2 = 63$; $\sum_i (y_i - \hat{y}_i)^2 = 15$. Оцените значимость уравнения в целом и коэффициента регрессии, определите коэффициент детерминации.

5. Фирма проводит рекламную компанию. Через семь недель фирма решила проанализировать эффективность этого вида рекламы, сопоставив недельные объёмы продаж (y, тыс. ден. ед.) с расходами на рекламу (x, тыс. ден. ед.):

$$\sum x = 48; \sum y = 526; \sum xy = 3694; \sum x^2 = 384; \sum y^2 = 39692.$$

Оценить параметры регрессии с помощью метода наименьших квадратов, предполагая, что зависимость y от x описывается уравнением $y = a + b \cdot x$.

6. На основании 10 наблюдений за пациентами, страдающими эмфиземой лёгких, врач-исследователь выясняет зависимость площади поражённой части лёгких (переменная y, измеряется в процентах от общей площади лёгких) от числа лет курения (переменная x, измеряется в годах). Получено уравнение регрессии:

$$\hat{y} = \underset{(5,6)}{11{,}2} + \underset{(0,4)}{1{,}3} \cdot x$$

В скобках под оценками параметров приведены их стандартные ошибки.

Задание:

а) Постройте 90% доверительный интервал для свободного члена.

б) Определите стаж курения, который приводит к поражению лёгких в среднем на 40%.

в) Проверьте гипотезу о том, что каждый дополнительный год курения приводит к увеличению поражённой площади в среднем на два процента, используя уровень значимости 5%.

7. По результатам исследования 15 предприятий было получено уравнение регрессии, выражающее зависимость объёма продаж *y* (тыс. ден. ед.) от расходов на рекламу x (тыс. ден. ед.): $\hat{y} = 12 + 0{,}6 \cdot x$. Средние квадратические отклонения $\sigma_x = 3{,}8$; $\sigma_y = 3{,}2$.

Задание:

1) Определить коэффициент корреляции.

2) Оценить значимость уравнения в целом.

3) Оценить значимость коэффициента регрессии через *t*-критерий Стьюдента.

8. По 20 наблюдениям построена следующая модель регрессии

$\hat{y} = 9 - 6 \cdot x$. Известно также, что, $\sum(x - \bar{x})^2 = 250, \sum(y - \hat{y}_x)^2 = 44$. Постройте доверительный интервал для коэффициента регрессии в этой модели с вероятностью 90%, 95% и 99%.

9. При исследовании зависимости объема производства *y* (тыс. ед.) от численности занятых x (чел.) по 25 предприятиям получено, что доля остаточной дисперсии в общей составляет 20%. Оценить значимость коэффициента корреляции.

10. По 21 субъекту РФ получена зависимость *y* – фактического конечного потребления домашних хозяйств на территории субъектов Российской Федерации (в текущих рыночных ценах; миллионов рублей) от x – валового регионального продукта по субъектам Российской Федерации (в текущих рыночных ценах; миллионов рублей)

$$y = 1{,}6225 \cdot x^{0{,}9244};\ \rho = 0{,}9.$$

Что означает параметр 0,9244? Какая доля дисперсии объясняется уравнением регрессии? Сколько приходится на не включённые в модель факторы?

11. Информация о деятельности 10 магазинов торговой сети представлена в таблице 9.

Задание: построить поле корреляции; определить форму связи, подобрать подходящие уравнение связи между результатом и фактором.

Таблица 9

№ магазина	Годовой товарооборот y, млн.руб.	Торговая площадь x, тыс.м^2
1	139	21
2	150	25
3	125	28
4	154	29
5	148	32
6	141	34
7	162	37
8	159	38
9	168	42
10	195	56

12. По исходным данным предыдущей задачи получены уравнения регрессии:

$y = 107{,}89 \cdot e^{0{,}0102 \cdot x} + \varepsilon;\ \rho^2 = 0{,}77.$

$y = 139{,}02 - 0{,}6823 \cdot x + 0{,}0305 \cdot x^2 + \varepsilon;\ \ \rho^2 = 0{,}79.$

$\ln y = 3{,}81 + 0{,}35 \cdot \ln x + \varepsilon;\ \ \rho^2 = 0{,}65.$

Задание:

1) оцените значимость каждого уравнения;

2) запишите функцию, характеризующую зависимость y от x в 3-м уравнении;

3) найдите коэффициенты эластичности для каждого уравнения при значении $x = 40$.

13. Моделирование прибыли фирмы привело к результатам

Таблица 10

Прибыль фирмы, тыс. руб., y											
	фактическая	10	12	15	17	18	11	12	9	11	9
	расчетная	11	11	17	15	20	12	13	10	10	9

Задание:

Найдите показатель тесноты связи прибыли с исследуемым в модели фактором. Оцените качество модели, определив среднюю ошибку аппроксимации и критерий Фишера, если $\hat{y} = a + b \cdot x + c \cdot x^2$.

14. В таблице 11 представлены данные по Дальневосточному федеральному округу за 2021 год.

Задание:

1) Постройте поле корреляции и сформулируйте гипотезу о форме связи.

2) Оцените с помощью метода наименьших квадратов параметры линейного регрессионного уравнения, Рассчитайте параметры уравнения парной регрессии и оцените его качество.

Таблица 11

Субъект РФ	Численность занятых, приходящихся на одного пенсионера (в среднем за год) *x*, человек	Средний размер назначенных пенсий *y*, руб.
Республика Бурятия	1,43	14 703
Республика Саха (Якутия)	1,81	20 793
Забайкальский край	1,67	14 679
Камчатский край	1,93	23 317
Приморский край	1,8	16 034
Хабаровский край	1,85	18 101
Амурская область	1,72	15 902
Магаданская область	2,12	23 455
Сахалинская область	1,69	20 660
Еврейская автономная область	1,41	15 345
Чукотский автономный округ	2,37	26 381

15. В таблице 12 представлены данные по Российской Федерации

Таблица 12

Год	2013	2014	2015	2016	2017	2018	2019	2020
Среднедушевой доход, руб.	25684	27412	30254	30865	31897	33178	35249	35142
Минимальный размер оплаты труда, руб.	5205	5554	5965	6204	7800	9489	11280	12130

Задание: построить поле корреляции; определить форму связи, подобрать подходящие уравнение связи между результатом и фактором.

16. В таблице 13 представлены данные по Дальневосточному федеральному округу за 2021 год.

Задание:

1) Постройте поле корреляции и сформулируйте гипотезу о форме связи.

Таблица 13

Субъект РФ	Среднедушевые денежные доходы населения *x*, руб./мес.	Потребительские расходы в среднем на душу населения *y*, руб./мес.
Республика Бурятия	28 314	24 835
Республика Саха (Якутия)	50 369	37 375
Забайкальский край	29 833	22 703
Камчатский край	60 791	36 968
Приморский край	40 843	33 695
Хабаровский край	44 101	38 515
Амурская область	39 626	30 865
Магаданская область	80 979	40 022
Сахалинская область	63 854	46 906
Еврейская автономная область	30 297	23 702
Чукотский автономный округ	99 912	32 846

2) С помощью MS Excel рассчитайте параметры уравнения парной линейной регрессии и оцените его качество. Постройте ещё одно уравнение регрессии, исключив из рассмотрения Чукотский автономный округ. Сравните полученные уравнения. Сделайте выводы.

17. Некоторая организация желает исследовать зависимость полученной прибыли *y* (сотни тыс. руб.) от вложения средств в научные разработки выпускаемой продукции *x* (тыс. руб.). Для этого рассматриваются 4 регрессионных уравнения:

линейное: $\hat{y} = a + b \cdot x$,

гиперболическое $\hat{y} = a + \frac{b}{x}$,

экспоненциальное $\hat{y} = e^{a + b \cdot x}$,

степенное $y = a \cdot x^b$.

В результате наблюдений, получены данные:

Таблица 14

Прибыль *y*	5	6	8	11	16	22	29	35	44	57	83
Вложения *x*	2	4	7	9	10	12	15	16	20	22	25

Задание:

1) подобрать уравнение, наилучшим образом описывающее данную зависимость;

2) изобразить регрессионные зависимости графически.

18. Изучалась зависимость вида $\hat{y} = a \cdot x^b$. Для преобразованных в логарифмах переменных получены следующие данные:

$$\sum XY = 106{,}3233\,;\ \sum X = 37{,}989\,;\ \sum X^2 = 207{,}619\,;$$

$$\sum Y = 20{,}038;\ \sum \left(Y - \hat{Y}_X\right)^2 = 0{,}516;\ n = 7\,;\ \sigma_y = 17{,}42.$$

Определите параметр b. Найдите показатель корреляции. Оцените его значимость. Оцените значимость уравнения регрессии.

19. По 12 областям имеются следующие данные о зависимости среднего размера ежемесячных назначенных пенсий y, тыс. руб. от прожиточного минимума в среднем на одного пенсионера в месяц x, тыс. руб.:

$$\sum x^2 = 491762\,;\ \sum (y - \hat{y}_x)^2 = 716{,}91;\ \sigma_x = 33{,}37\,;\ \sigma_y = 9{,}67.$$

Оценить значимость параметров линейной регрессии и корреляции.

20. Для трёх видов продукции A, B, C получены модели зависимости удельных постоянных расходов от объёма выпускаемой продукции:

$$\hat{y}_A = 280 + 0{,}6 \cdot x,$$

$$\hat{y}_B = 80 + 1{,}4 \cdot \sqrt{x},$$

$$\hat{y}_C = 20 \cdot x^{0{,}6}.$$

Задание:

1) Определить коэффициенты эластичности по каждому виду продукции и пояснить их смысл.
2) Сравнить эластичности затрат при объёме выпуска продукции равном 1000.
3) При каком объёме выпускаемой продукции коэффициенты эластичности для продукции A и C будут равны.

21. Зависимость спроса на некоторый товар от его цены характеризуется уравнением $\lg y = 1{,}55 - 0{,}45 \cdot \lg x$. Доля остаточной дисперсии в общей составляет $0{,}2$, $n = 25$.

Задание:

1) Запишите данное уравнение в степенной форме.
2) Найдите индекс корреляции.
3) Оцените значимость уравнения регрессии.

Домашнее задание к разделу парная регрессия

Задача 1

Для проверки эффективности рекламы товара, фирма решила провести эксперимент в 10 регионах, имеющих идентичные средние показатели продаж. В рамках эксперимента была введена различная рекламная стратегия, с выделением бюджета *x* для рекламных мероприятий. При этом фиксировалось число продаж *y*.

Задание:

1. Постройте поле корреляции и сформулируйте гипотезу о форме связи.

2. Постройте уравнение линейной регрессии.

3. Рассчитайте коэффициент линейной корреляции и коэффициент детерминации.

4. Оцените с помощью средней ошибки аппроксимации качество уравнения.

5. Дайте с помощью среднего коэффициента эластичности сравнительную оценку силы связи фактора с результатом.

6. С помощью *t*-критерия Стьюдента оцените статистическую значимость параметров регрессии (найдите их доверительные интервалы) и коэффициента корреляции.

7. Оцените с помощью *F*-критерия Фишера значимость уравнения.

8. Рассчитайте прогнозное значение результата, если прогнозное значение фактора увеличится на 10% от его среднего уровня. Оцените точность прогноза, рассчитав ошибку прогноза и его доверительный интервал.

9. Расчёты выполните вручную и с помощью MS Excel. Сравните полученные результаты. Оформите выполненное задание в виде отчета.

Таблица 15

Вариант	Расходы на рекламу x_i, млн. р. (одинаковое для всех вариантов)									
	0	0,5	1	1,5	2	2,5	3	3,5	4	4,5
	Количества продаж y_i, тыс. ед. (по вариантам)									
1	12,3	16,3	16,4	16,0	18,5	17,3	20,0	19,5	19,0	19,7
2	39,5	40,3	40,7	40,8	43,1	42,7	45,3	46,2	47,4	49,5
3	32,4	32,4	34,8	37,1	38,0	38,7	38,6	39,9	43,8	43,5
4	21,0	23,0	23,7	23,8	25,8	27,6	28,4	29,7	31,7	31,6
5	27,6	28,8	29,6	31,1	30,9	31,3	33,1	34,6	35,1	37,2

Вариант	Расходы на рекламу x_i, млн. р. (одинаковое для всех вариантов)									
	0	0,5	1	1,5	2	2,5	3	3,5	4	4,5
	Количества продаж y_i, тыс. ед. (по вариантам)									
6	30,6	32,8	32,1	33,7	35,1	39,2	37,4	39,7	42,3	43,4
7	18,5	19,5	20,1	23,7	23,6	24,0	26,2	26,5	28,3	28,1
8	13,3	12,2	13,1	11,5	15,7	13,7	16,8	13,9	16,9	16,8
9	14,2	16,3	16,6	18,9	19,4	20,4	23,3	24,2	27,1	27,4
10	34,4	34,8	36,1	37,7	37,3	37,5	37,5	39,6	40,9	43,6
11	20,6	20,2	19,6	21,3	23,2	23,9	23,2	23,0	24,1	25,2
12	17,4	18,6	18,0	21,3	21,3	24,4	24,1	27,2	27,0	28,7
13	38,3	39,3	40,1	43,9	42,9	42,1	45,2	44,3	47,9	47,8
14	38,0	40,9	39,1	39,7	39,3	38,4	41,4	42,9	41,3	42,7
15	36,7	36,5	37,2	38,0	38,3	39,5	41,7	39,9	42,0	41,8
16	38,1	38,6	40,9	38,6	41,3	43,1	44,3	43,0	45,8	46,2
17	30,8	31,1	30,4	31,7	30,5	33,5	31,0	34,5	36,0	32,9
18	10,7	11,0	13,2	12,4	13,2	13,3	14,4	15,3	14,8	14,8
19	23,7	24,8	25,8	27,6	26,9	25,2	26,6	26,3	29,0	30,4
20	22,8	26,3	28,0	26,1	26,0	29,9	30,9	32,9	33,9	33,5
21	26,5	26,4	28,2	26,7	29,1	29,7	29,7	31,2	32,1	32,4
22	25,3	28,8	30,1	30,0	32,5	31,4	32,0	36,4	35,6	36,9
23	10,0	9,7	11,6	12,2	13,3	13,9	15,6	16,7	15,1	16,8
24	20,9	20,7	20,8	20,9	22,8	22,4	24,5	22,9	22,7	24,6
25	24,8	26,5	28,3	29,1	27,0	28,4	30,0	32,4	32,0	32,3
26	29,4	30,0	32,0	33,1	32,6	33,9	33,6	35,0	34,7	35,9
27	20,3	20,4	22,1	24,3	25,1	25,1	26,9	25,4	27,8	26,9
28	20,8	20,2	21,5	21,8	24,4	23,7	25,7	24,7	27,2	24,8
29	28,6	28,6	28,8	29,2	31,7	32,7	32,1	33,3	33,8	35,0
30	16,1	17,0	20,5	17,1	18,8	21,0	22,7	24,2	23,4	26,7

Задача 2

Имеются данные о доли расходов на товары длительного пользования y от среднемесячного дохода семьи x.

Задание:

1. Постройте поле корреляции и сформулируйте гипотезу о форме связи.

2. Постройте уравнение нелинейной регрессии, предполагая, что эта зависимость выражается степенной функцией $y = a \cdot x^b$.

3. Рассчитайте показатель корреляции, детерминации.

4. Найдите среднюю ошибку аппроксимации.

5. Оцените с помощью F-критерия Фишера значимость уравнения.

Таблица 16

Вариант	Доход семьи x_i, тыс.р. на 1 чел. (для всех вариантов)									
	2	3,5	4	5	5,5	6,5	8	9	11	14
	Процент расходов на товары длительного пользования y_i									
1	29,3	25,4	25,0	23,4	23,1	22,6	21,7	21,7	22,2	22,4
2	31,2	27,0	26,1	26,1	23,1	23,8	22,3	21,4	21,8	22,5
3	29,7	26,3	24,8	23,5	22,3	21,7	21,5	19,0	20,5	22,8
4	20,4	19,7	16,6	17,3	15,1	15,2	14,3	14,1	14,3	14,1
5	30,7	27,0	25,1	24,1	21,3	22,7	23,7	20,8	19,8	21,9
6	29,7	28,2	24,6	24,6	22,8	22,2	22,0	21,8	23,3	21,5
7	31,4	28,4	27,3	24,9	23,5	23,6	23,2	21,8	23,3	22,1
8	27,9	25,4	20,7	23,6	21,6	20,1	21,3	21,2	20,8	18,5
9	27,0	23,4	22,1	20,5	19,3	18,9	17,3	16,7	17,7	16,1
10	30,0	27,9	25,7	23,7	21,8	21,7	22,0	19,3	22,2	19,5
11	29,5	27,2	23,4	21,9	21,3	22,2	21,0	20,0	20,2	19,6
12	29,8	26,9	24,3	23,7	23,0	23,2	20,7	21,9	21,0	20,7
13	26,7	24,5	19,5	21,5	21,0	18,0	16,5	16,2	17,2	17,8
14	24,7	21,5	22,1	21,9	20,3	19,1	20,6	20,2	18,7	20,3
15	27,1	23,9	25,1	20,9	21,6	20,6	20,5	19,1	21,8	20,6
16	27,9	24,3	22,1	21,8	20,7	17,9	17,8	19,5	15,8	20,1
17	23,2	19,7	19,2	16,5	16,7	17,8	16,2	16,8	14,5	15,6
18	23,1	22,4	19,1	18,3	16,7	15,3	17,3	16,2	14,7	15,8
19	27,8	25,3	25,2	24,9	24,7	24,8	23,4	22,9	21,4	22,0
20	19,9	19,4	17,5	17,2	16,5	16,1	13,5	13,8	15,1	13,2
21	25,1	21,9	21,9	19,7	17,9	18,0	18,7	17,5	16,5	16,2

Вариант	Доход семьи x_i, тыс.р. на 1 чел. (для всех вариантов)									
	2	3,5	4	5	5,5	6,5	8	9	11	14
	Процент расходов на товары длительного пользования y_i									
22	27,7	27,6	26,4	24,7	24,5	23,9	23,9	22,6	23,7	21,7
23	23,0	21,7	20,6	20,3	19,6	16,9	19,1	18,9	16,0	16,4
24	25,5	23,4	21,6	19,7	18,3	17,6	18,3	16,9	18,0	18,2
25	20,4	16,9	16,7	16,8	15,6	14,9	12,7	12,0	14,2	13,5
26	32,6	31,1	25,8	24,7	25,6	24,7	22,9	24,5	22,7	22,5
27	20,8	19,9	19,0	18,6	17,7	16,9	18,3	15,8	14,2	14,3
28	19,3	17,8	15,4	16,0	15,5	14,5	15,2	15,3	13,1	14,1
29	26,1	20,5	20,9	18,7	18,4	18,5	17,4	18,5	13,7	15,8
30	27,1	24,4	22,2	20,9	20,4	18,3	19,0	19,4	20,0	19,6

МНОЖЕСТВЕННАЯ РЕГРЕССИЯ

Понятие множественной регрессии

Множественной регрессией называют модель, выражающую зависимость среднего значения зависимой переменной y от нескольких независимых переменных $x_1, x_2, \ldots, x_p$

$$\hat{y} = f(x_1,\ x_2,\ \ldots,\ x_p).$$

Множественная регрессия применяется в ситуациях, когда из множества факторов, влияющих на результативный признак, нельзя выделить один доминирующий фактор и необходимо учитывать одновременное влияние нескольких факторов.

Постановка задачи

По имеющимся данным n наблюдений за совместным изменением $p+1$ переменной y и x_j $\{(x_{ji}, y_i), j = 1, 2, \ldots, p;\ i = 1, 2, \ldots, n\}$ необходимо определить аналитическую зависимость $\hat{y} = f(x_1,\ x_2,\ \ldots,\ x_p)$, наилучшим образом описывающую данные наблюдений.

Таблица 17

№	y	x_1	x_2	…	x_p
1	y_1	x_{11}	x_{21}	…	x_{p1}
2	y_2	x_{12}	x_{22}	…	x_{p2}
…	…	…	…	…	…
n	y_n	x_{1n}	x_{2n}	…	x_{pn}

Отбор факторов

Применительно к множественной регрессии, необходимо до определения вида модели, произвести отбор факторов. Факторы, включаемые в модель:

1) должны быть количественно измеримы;

2) должны существенно влиять на вариацию независимой переменной;

3) не должны быть взаимно коррелированы и, тем более, находиться в точной функциональной связи.

Если $r_{x_i x_j} \geq 0{,}7$, то x_i и x_j коллинеарны, или находятся в линейной зависимости между собой.

Напомним, что коэффициент линейной корреляции определяется по формуле:

$$r_{x_i x_j} = \frac{\overline{x_j \cdot x_i} - \bar{x}_j \cdot \bar{x}_i}{\sigma_{x_i} \cdot \sigma_{x_j}}.$$

Разновидностью интеркоррелированности факторов является мультиколлинеарность – наличие высокой линейной связи между всеми или несколькими факторами.

Если между факторами есть полная линейная зависимость и все коэффициенты корреляции равны 1, то

$$|R| = \begin{vmatrix} 1 & 1 & \ldots & 1 \\ 1 & 1 & \ldots & 1 \\ \ldots & \ldots & \ldots & \ldots \\ 1 & 1 & \ldots & 1 \end{vmatrix} = 0.$$

Мультиколлинеарность имеет место, если определитель матрицы межфакторной корреляции близок к нулю.

Если же определитель матрицы межфакторной корреляции близок к единице, то мультколлинеарности нет.

$$|R| = \begin{vmatrix} r_{x_1 x_1} & r_{x_2 x_1} & \cdots & r_{x_1 x_p} \\ r_{x_2 x_1} & r_{x_2 x_2} & \cdots & r_{x_2 x_p} \\ \cdots & \cdots & \cdots & \cdots \\ r_{x_p x_1} & r_{x_2 x_2} & \ldots & r_{x_p x_p} \end{vmatrix} = \begin{vmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \ldots & 1 \end{vmatrix} = 1.$$

Линейная модель множественной регрессии

Линейная множественная регрессия описывается уравнением:

$$\hat{y} = a + b_1 x_1 + \cdots + b_p x_p \text{ или } y = a + b_1 x_1 + \cdots + b_p x_p + \varepsilon.$$

В линейной множественной регрессии параметры b_j при x называются *коэффициентами чистой регрессии*, они характеризуют среднее изменение результата с изменением соответствующего фактора на 1 единицу при неизменном значении других факторов, закрепленных на среднем значении.

Нелинейные модели множественной регрессии

Для построения уравнения регрессии чаще всего используются функции:

$y = a \cdot x_1^{b_1} \cdot x_2^{b_2} \cdot \ldots \cdot x_p^{b_p} \cdot \varepsilon$ – степенная,

показатели степеней b_j являются коэффициентами эластичности. Они показывают, на сколько процентов изменяется в среднем результат с изменением соответствующего фактора на 1% при неизменности действия других факторов.

Этот вид уравнения регрессии получил наибольшее распространение в производственных функциях, в исследованиях спроса и потребления. В частности, производственная функция Кобба – Дугласа имеет вид:

$$y = a \cdot x_1^{b_1} \cdot x_2^{b_2},$$

где x_1 – капиталовложения, x_2 – трудозатраты, а y – совокупный выпуск продукции,

$y = e^{a+b_1x_1+\cdots+b_px_p+\varepsilon}$ – экспоненциальная,

$y = \frac{1}{a+b_1x_1+\cdots+b_px_p+\varepsilon}$ – гиперболическая.

Исследователь может использовать любые другие функции, которые после соответствующих преобразований приводятся к линейному виду.

Оценка параметров уравнения множественной регрессии

Для оценки параметров уравнения множественной регрессии применяется МНК. Для линейных уравнений (и нелинейных уравнений, приводимых к линейным) строится система нормальных уравнений.

$$\begin{cases} \sum y = n \cdot a + b_1 \cdot \sum x_1 + b_2 \cdot \sum x_2 + \cdots + b_p \cdot \sum x_p\,, \\ \sum yx_1 = a \cdot \sum x_1 + b_1 \cdot \sum x_1^2 + b_2 \cdot \sum x_2 \cdot x_1 + \cdots + b_p \cdot \sum x_p \cdot x_1\,, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \sum yx_p = a \cdot \sum x_p + b_1 \cdot \sum x_1 \cdot x_p + b_2 \cdot \sum x_2 \cdot x_p + \cdots + b_p \cdot \sum x_p^2. \end{cases}$$

Решение данной системы позволяет получить оценки параметров.

Уравнение регрессии в стандартизованном масштабе

Возможен иной подход к определению параметров, когда на основе матрицы парных коэффициентов корреляции строится уравнение регрессии в стандартизованном масштабе

$$t_y = \beta_1 t_{x_1} + \beta_2 t_{x_1} + \cdots + \beta_p t_{x_p} + \varepsilon,$$

$$t_y = \frac{y - \bar{y}}{\sigma_y}, \quad t_{x_j} = \frac{x_j - \bar{x}_j}{\sigma_{x_j}}.$$

t_y и t_{x_j} – стандартизованные переменные, для которых среднее значение равно нулю $\bar{t}_y = \bar{t}_{x_j} = 0$, а среднее квадратичное отклонение равно единице $\sigma_y = \sigma_{x_j} = 1$.

Величины β_j называются *стандартизованными коэффициентами регрессии*.

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

$$\begin{cases} r_{yx_1} = \beta_1 + \beta_2 r_{x_1x_2} + \cdots + \beta_p r_{x_1x_p}, \\ r_{yx_2} = \beta_1 r_{x_2x_1} + \beta_2 + \cdots + \beta_p r_{x_2x_p}, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ r_{yx_p} = \beta_1 r_{x_px_1} + \beta_2 r_{x_px_2} + \cdots + \beta_p. \end{cases}$$

Стандартизированные коэффициенты определяются из решения данной системы.

Для случая двух переменных параметры уравнения

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

можно найти по формулам

$$\beta_1 = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1x_2}}{1 - r_{x_1x_2}^2}, \qquad \beta_2 = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1x_2}}{1 - r_{x_1x_2}^2}.$$

Коэффициенты b_j линейной множественной регрессии в исходных (натуральных) показателях связаны со стандартизованными коэффициентами β_j соотношением

$$b_j = \beta_j \frac{\sigma_y}{\sigma_{x_j}}.$$

Свободный член a уравнения в натуральных показателях определяется как $a = \bar{y} - b_1\bar{x}_1 - \cdots - b_p\bar{x}_p$.

Стандартизованные параметры β_j показывают, на сколько σ_y изменится в среднем результат, если увеличить соответствующий фактор x_j на одну σ_{x_j} при неизменном среднем уровне других факторов.

Основное достоинство стандартизованных коэффициентов регрессии заключается в том, что они сравнимы между собой. Сравнивая их, можно ранжировать факторы по силе их воздействия на результат. Коэффициенты «чистой» регрессии несравнимы между собой.

Коэффициент эластичности

Средние коэффициенты эластичности для линейной множественной регрессии

$$\bar{э}_{yx_j} = b_j \frac{\bar{x}_j}{\bar{y}}$$

Показывают на сколько процентов в среднем по совокупности изменится результат *y* от своей величины при изменении фактора *x* на 1% от своего значения при неизменных значениях других факторов.

Средние показатели эластичности можно сравнивать друг с другом и соответственно ранжировать факторы по силе их воздействия на результат.

Оценка тесноты связи

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*

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

где $\sigma^2_y = \frac{1}{n}\sum_i (y_i - \bar{y})^2$ – общая дисперсия результативного признака *y*,

$\sigma^2_{ост} = \frac{1}{n}\sum_i (y_i - \hat{y}_i)^2$ – остаточная дисперсия для уравнения $\hat{y} = f(x_1,\ x_2,\ \ldots,\ x_p)$.

Индекс множественной корреляции принимает значения от 0 до 1. Чем ближе его значение к 1, тем теснее связь результативного признака со всем набором исследуемых факторов. Его величина должна быть больше или равна максимального парного индекса корреляции

$$R_{yx_1x_2\ldots x_p} \geq \max_j r_{yx_j}\ (j = 1, \ldots, p).$$

Для уравнения в стандартизированном масштабе индекс множественной корреляции можно выразить через стандартизованные коэффициенты и парные коэффициенты линейной корреляции

$$R_{yx_1x_2\ldots x_p} = \sqrt{\sum_j \beta_j r_{yx_j}}.$$

При линейной зависимости *коэффициент множественной корреляции* можно определить через матрицу парных коэффициентов корреляции

$$R_{yx_1x_2\ldots x_p} = \sqrt{1 - \frac{\Delta_r}{\Delta_{r_{11}}}},$$

где Δ_r – определитель матрицы парных коэффициентов корреляции

$$\Delta_r = \begin{vmatrix} 1 & r_{yx_1} & \ldots & r_{yx_p} \\ r_{yx_1} & 1 & \ldots & r_{x_1x_p} \\ \ldots & \ldots & \ldots & \ldots \\ r_{yx_p} & r_{x_px_1} & \ldots & 1 \end{vmatrix},$$

а $\Delta_{r_{11}}$ – определитель матрицы межфакторной корреляции

$$\Delta_{r_{11}} = \begin{vmatrix} 1 & r_{x_1x_2} & \ldots & r_{x_1x_p} \\ r_{x_2x_1} & 1 & \ldots & r_{x_2x_p} \\ \ldots & \ldots & \ldots & \ldots \\ r_{x_px_1} & r_{x_px_2} & \ldots & 1 \end{vmatrix}.$$

Частная корреляция

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

$$r_{yx_i*x_1x_2\ldots x_{i-1}x_{i+1}\ldots x_p} = \sqrt{1 - \frac{1 - R^2_{yx_1x_2\ldots x_i\ldots x_p}}{1 - R^2_{yx_1x_2\ldots x_{i-1}x_{i+1}\ldots x_p}}},$$

где $R^2_{yx_1x_2\ldots x_i\ldots x_p}$ – множественный коэффициент детерминации всего комплекса p факторов с результатом,

$R^2_{yx_1x_2\ldots x_{i-1}x_{i+1}\ldots x_p}$ – показатель детерминации, но без ввведения в модель фактора x_i.

Для двухфакторной модели формула принимает вид

$$r_{yx_1*x_2} = \sqrt{1 - \frac{1 - R^2_{yx_1x_2}}{1 - r^2_{yx_2}}}, \qquad r_{yx_2*x_1} = \sqrt{1 - \frac{1 - R^2_{yx_1x_2}}{1 - r^2_{yx_1}}}.$$

Рекурентная формула

$$r_{yx_i*x_1x_2\ldots x_{i-1}x_{i+1}\ldots x_p} = \frac{r_{yx_i*x_1x_2\ldots x_{i-1}x_{i+1}\ldots x_{p-1}} - r_{yx_p*x_1x_2\ldots x_{p-1}} \cdot r_{x_ix_p*x_1x_2\ldots x_{i-1}x_{i+1}\ldots x_{p-1}}}{\sqrt{\left(1 - r^2_{yx_p*x_1x_2\ldots x_{p-1}}\right) \cdot \left(1 - r^2_{x_ix_p*x_1x_2\ldots x_{i-1}x_{i+1}\ldots x_{p-1}}\right)}}.$$

Для двухфакторной модели формула принимает вид

$$r_{yx_1*x_2} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1x_2}}{\sqrt{\left(1 - r^2_{yx_2}\right) \cdot \left(1 - r^2_{x_1x_2}\right)}}, \qquad r_{yx_2*x_1} = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1x_2}}{\sqrt{\left(1 - r^2_{yx_1}\right) \cdot \left(1 - r^2_{x_1x_2}\right)}}.$$

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 подтверждают ранжировку факторов по силе воздействия на результат, полученную на основе стандартизированных коэффициентов регрессии β_j, но, в отличие от последних, дают меру тесноты связи каждого фактора с результатом в чистом виде.

Оценка качества построенной модели

Качество построенной модели в целом оценивает коэффициент (индекс) детерминации.

Коэффициент (индекс) множественной детерминации рассчитывается как квадрат коэффициента (индекса) множественной корреляции $R^2_{yx_1x_2\ldots x_p}$.

Коэффициент детерминации характеризует долю дисперсии результативного признака *y*, объясняемую регрессией, в общей дисперсии результативного признака

Скорректированный коэффициент (индекс) множественной детерминации содержит поправку на число степеней свободы и рассчитывается по формуле

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-m-1},$$

где n – число наблюдений, m – число параметров при факторных переменных в уравнении множественной регрессии.

Оценка значимости уравнения регрессии в целом и его параметров

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

$$F = \frac{S^2_{факт}}{S^2_{ост}} = \frac{R^2}{1-R^2}\frac{n-m-1}{m}.$$

R^2 – коэффициент (индекс) множественной детерминации

Частный F-критерий позволяет оценить статистическую значимость вклада в уравнение регрессии фактора x_i.

$$F_{x_i} = \frac{R^2_{yx_1x_2\ldots x_i\ldots x_p} - R^2_{yx_1x_2\ldots x_{i-1}x_{i+1}\ldots x_p}}{1 - R^2_{yx_1x_2\ldots x_i\ldots x_p}} \cdot \frac{n-m-1}{1}.$$

Фактическое значение частного *F*-критерия сравнивается с табличным $F_{\text{таб}}(\alpha; k_1; k_2)$ при уровне значимости α и числе степеней свободы: $k_1 = 1$ и $k_2 = n - m - 1$. Если фактическое значение F_{x_i} превышает $F_{\text{таб}}$, 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:

$$F_{x_1} = \frac{R^2_{yx_1x_2} - r^2_{yx_2}}{1 - R^2_{yx_1x_2}} \cdot \frac{n-m-1}{1}, \qquad F_{x_2} = \frac{R^2_{yx_1x_2} - r^2_{yx_1}}{1 - R^2_{yx_1x_2}} \cdot \frac{n-m-1}{1}.$$

Оценка значимости коэффициентов чистой регрессии по *t*-критерию Стьюдента сводится к вычислению значений:

$$t_{b_i} = \frac{b_i}{m_{b_i}} = \sqrt{F_{x_i}},$$

где m_{b_i} – средняя квадратическая ошибка коэффициента b_i, которая может быть определена по формуле

$$m_{b_i} = \frac{\sigma_y \sqrt{1 - R^2_{yx_1\ldots x_p}}}{\sigma_{x_i} \sqrt{1 - R^2_{x_i x_1\ldots x_p}}} \frac{1}{\sqrt{n-m-1}}.$$

σ_y – средне квадратическое отклонение для *y*;

σ_{x_i} – средне квадратическое отклонение для x_i;

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

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

Зная величину F_{x_i}, можно определить и *t*-критерий для коэффициента регрессии при *i* -м факторе, t_{b_i}, а именно:

$$t_{b_i} = \sqrt{F_{x_i}}.$$

Фактическое значение t-критерия сравнивается с табличным при уровне значимости α и $n-m-1$ степенях свободы. Если фактическое значение t_{b_i} превышает $t_{\text{таб}}(\alpha;\, n-m-1)$, то коэффициент чистой регрессии b_i при факторе x_i статистически значим. В противном случае – незначим.

Регрессия с фиктивными переменными

Иногда в процессе эконометрического моделирования у исследователя возникает потребность учитывать в качестве объясняющих факторов не только количественные, но и качественные характеристики. Чтобы ввести такие переменные в регрессионную модель, им должны быть присвоены те или иные цифровые метки. В этом случаях удобно использовать так называемые фиктивные переменные.

Фиктивная переменная – сконструированная количественная переменная, описывающая качественные факторы (например, пол, профессия, образование, принадлежность к какой-либо группе).

Часто такие переменные принимают одно из двух значений – 0 или 1. Их также называют бинарными или дамми-переменными (dummy variable). Например, в результате опроса группы людей 0 может означать, что опрашиваемый – мужчина, а 1 – женщина.

На практике количество фиктивных переменных в модели на 1 меньше, чем число градаций признака.

Решение типовых задач

Задача 1. Изучается зависимость прибыли y (тыс. руб.) от выработки продукции на одного работника x_1 (ед.) и индекса цен на продукцию x_2 (%).

Средние значения: $\bar{y} = 150, \bar{x}_1 = 35, \bar{x}_2 = 110$.

Средние квадратические отклонения $\sigma_y = 18, \sigma_{x_1} = 10, \sigma_{x_2} = 17$.

Парные коэффициенты корреляции $r_{yx_1} = 0{,}69,\ r_{yx_2} = 0{,}63;\ r_{x_1x_2} = 0{,}4$.

Число наблюдений $n = 25$.

Задание:

1) Найти уравнение множественной регрессии в стандартизированном масштабе. Дать интерпретацию коэффициентам.
2) Найти линейное уравнение множественной регрессии. Дать интерпретацию коэффициентам.
3) Рассчитать множественный и частные коэффициенты корреляции.
4) Рассчитать общий и частные критерии Фишера, сделать выводы.

Решение

1) стандартизированные коэффициенты β_j определим по формулам

$$\beta_1 = \frac{r_{yx_1} - r_{yx_2}r_{x_1x_2}}{1 - r^2_{x_1x_2}} = \frac{0{,}69 - 0{,}63 \cdot 0{,}4}{1 - 0{,}4^2} = 0{,}52.$$

$$\beta_2 = \frac{r_{yx_2} - r_{yx_1}r_{x_1x_2}}{1 - r^2_{x_1x_2}} = \frac{0{,}63 - 0{,}69 \cdot 0{,}4}{1 - 0{,}4^2} = 0{,}42.$$

Уравнение регрессии в стандартизированном виде

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

$$t_y = 0{,}52 t_{x_1} + 0{,}42 t_{x_2}.$$

Так как $\beta_1 > \beta_2$, то выработки продукции на одного работника оказывает на прибыль большее влияние, чем индекс цен на продукцию.

2) определим коэффициенты чистой регрессии через стандартизированные

$$b_j = \beta_j \frac{\sigma_y}{\sigma_{x_j}}.$$

$$b_1 = 0{,}52 \cdot \frac{18}{10} = 0{,}936.$$

$$b_2 = 0{,}42 \cdot \frac{18}{17} = 0{,}445.$$

Свободный член определим по формуле

$$a = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2$$

$$a = 150 - 0{,}936 \cdot 35 - 0{,}445 \cdot 110 = 68{,}29.$$

Уравнение множественной регрессии

$$\hat{y} = 68{,}29 + 0{,}936 \cdot x_1 + 0{,}445 \cdot x_2$$

Если выработка продукции на одного работника увеличится на 1 единицу, то прибыль в среднем увеличится на 0,936 тыс. руб. (или 936 рублей) при неизменном уровне индекса цен на продукцию.

Если индекс цен на продукцию увеличится на 1%, то прибыль в среднем увеличится на 0,445 тыс. руб. (или 445 рублей) при неизменном уровне выработки продукции на одного работника.

3) Множественный коэффициент корреляции рассчитаем по формуле

$$R_{yx_1x_2} = \sqrt{\beta_1 r_{yx_1} + \beta_2 r_{yx_2}} = \sqrt{0{,}52 \cdot 0{,}69 + 0{,}42 \cdot 0{,}63} = \sqrt{0{,}6108} = 0{,}78.$$

Частные коэффициенты корреляции

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1x_2}}{\sqrt{\left(1 - r_{yx_2}^2\right)\left(1 - r_{x_1x_2}^2\right)}} = \frac{0{,}69 - 0{,}63 \cdot 0{,}4}{\sqrt{(1 - 0{,}63^2)(1 - 0{,}4^2)}} = 0{,}62.$$

$$r_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} r_{x_1x_2}}{\sqrt{\left(1 - r_{yx_1}^2\right)\left(1 - r_{x_1x_2}^2\right)}} = \frac{0{,}63 - 0{,}69 \cdot 0{,}4}{\sqrt{(1 - 0{,}69^2)(1 - 0{,}4^2)}} = 0{,}53.$$

Значения частных коэффициентов корреляции немного меньше парных $r_{yx_1 \cdot x_2} = 0{,}62 < r_{yx_1} = 0{,}69$ и $r_{yx_2 \cdot x_1} = 0{,}53 < r_{yx_2} = 0{,}63$ (отличаются из-за межфакторной связи, близкой к умеренной $r_{x_1x_2} = 0{,}4$), но выводы о направлении и тесноте связи совпадают. Связь прямая, умеренная, *y* с x_1 связан сильнее чем связь с x_2.

4) *F*-критерия Фишера для уравнения множественной регрессии

$$F_{\text{факт}} = \frac{R^2}{1 - R^2} \cdot \frac{n - m - 1}{m} = \frac{0{,}6108}{1 - 0{,}6108} \cdot \frac{25 - 2 - 1}{2} = 17{,}26.$$

Табличное значение критерия Фишера $F_{\text{таб}}(0{,}05; 2; 22) = 3{,}44$. Так как $F_{\text{факт}} > F_{\text{таб}}$, то уравнение в целом статистически значимо.

Частные критерии Фишера

$$F_{x_1} = \frac{R_{yx_1x_2}^2 - r_{yx_1}^2}{1 - R_{yx_1x_2}^2} \cdot (n - m - 1) = \frac{0{,}6108 - 0{,}4761}{1 - 0{,}6108} \cdot 22 = 7{,}61.$$

$$F_{x_2} = \frac{R^2_{yx_1x_2} - r^2_{yx_2}}{1 - R^2_{yx_1x_2}} \cdot (n - m - 1) = \frac{0{,}6108 - 0{,}3969}{1 - 0{,}6108} \cdot 22 = 12{,}09.$$

Табличное значение критерия Фишера $F_{\text{таб}}(0{,}05; 1; 22) = 4{,}30$.

$F_{x_1} > F_{\text{таб}}$, следовательно, включение фактора x_1 после фактора x_2 целесообразно.

$F_{x_2} > F_{\text{таб}}$, следовательно, включение фактора x_2 после фактора x_1 целесообразно.

Или можно сказать, что включение обоих факторов в модель целесообразно.

Задача 2. Найдите коэффициент множественной корреляции $R_{yx_1x_2}$, если коэффициенты парной корреляции равны: $r_{yx_1} = 0{,}78$, $r_{yx_2} = 0{,}81$, $r_{x_1x_2} = 0{,}33$.

Решение

Воспользуемся формулой $R_{yx_1x_2} = \sqrt{1 - \frac{\Delta_r}{\Delta_{r_{11}}}}$.

Для двухфакторной модели Δ_r является определителем 3-го порядка (найдём по правилу треугольников), а $\Delta_{r_{11}}$ – определителем 2-го порядка. Тогда

$$R_{yx_1x_2} = \sqrt{1 - \frac{1 + 2 \cdot r_{yx_1} \cdot r_{yx_2} \cdot r_{x_1x_2} - r^2_{yx_1} - r^2_{yx_2} - r^2_{x_1x_2}}{1 - r^2_{x_1x_2}}} =$$

$$= \sqrt{\frac{1 - r^2_{x_1x_2} - \left(1 + 2 \cdot r_{yx_1} \cdot r_{yx_2} \cdot r_{x_1x_2} - r^2_{yx_1} - r^2_{yx_2} - r^2_{x_1x_2}\right)}{1 - r^2_{x_1x_2}}} =$$

$$= \sqrt{\frac{r^2_{yx_1} + r^2_{yx_2} - 2 \cdot r_{yx_1} \cdot r_{yx_2} \cdot r_{x_1x_2}}{1 - r^2_{x_1x_2}}} =$$

$$= \sqrt{\frac{0{,}78^2 + 0{,}81^2 - 2 \cdot 0{,}78 \cdot 0{,}81 \cdot 0{,}33}{1 - 0{,}33^2}} = \sqrt{0{,}951} = 0{,}975.$$

Связь между y и факторами x_1, x_2 очень тесная.

Задача 3. Имеется информация по 25 наблюдениям:

средние значения: $\bar{y} = 35,\ \bar{x}_1 = 16,\ \bar{x}_2 = 8$;

коэффициенты вариации: $V_y = 20\%,\ V_{x_1} = 30\%,\ V_{x_2} = 10\%$;

уравнения регрессии: $\hat{y} = 18 + 1{,}1x_1 - 2x_2,\ \ \hat{y} = 8 + x_1,\ \ \hat{y} = 4 - 3{,}1x_2$.

Задание:

1) Оценить значимость уравнения множественной регрессии, если известно, что $r_{x1x2} = -0{,}25$.

2) Оценить значимость коэффициентов 1-го уравнения.

Решение

1) Значимость уравнения проверяется с помощью *F*-критерия Фишера. Для расчёта *F*-критерия нужен множественный коэффициент корреляции, который можно рассчитать через парные коэффициенты корреляции. Коэффициенты парной корреляции найдем по формуле

$$r_{xy} = b\frac{\sigma_x}{\sigma_y}.$$

Коэффициент b возьмём из соответствующего уравнения парной регрессии. Среднеквадратические отклонения выразим из коэффициентов вариации

$$V_y = \frac{\sigma_y}{\bar{y}} \cdot 100\% \to \sigma_y = \frac{V_y \cdot \bar{y}}{100\%} = \frac{20 \cdot 35}{100} = 7.$$

$$V_{x_1} = \frac{\sigma_{x_1}}{\overline{x_1}} \cdot 100\% \to \sigma_{x_1} = \frac{V_{x_1} \cdot \overline{x_1}}{100\%} = \frac{30 \cdot 16}{100} = 4{,}8.$$

$$V_{x_2} = \frac{\sigma_{x_2}}{\overline{x_2}} \cdot 100\% \to \sigma_{x_2} = \frac{V_{x_2} \cdot \overline{x_2}}{100\%} = \frac{10 \cdot 8}{100} = 0{,}8.$$

$$r_{yx_1} = b\frac{\sigma_{x_1}}{\sigma_y} = 1 \cdot \frac{4{,}8}{7} = 0{,}68.$$

$$r_{yx_2} = b\frac{\sigma_{x_2}}{\sigma_y} = -3{,}1\frac{0{,}8}{7} = -0{,}35.$$

Теперь найдём множественный коэффициент корреляции

$$R_{yx_1x_2} = \sqrt{\frac{r_{yx_1}^2 + r_{yx_2}^2 - 2 \cdot r_{yx_1} \cdot r_{yx_2} \cdot r_{x_1x_2}}{1 - r_{x_1x_2}^2}} =$$

$$= \sqrt{\frac{0{,}68^2 + (-0{,}35)^2 - 2 \cdot 0{,}68 \cdot (-0{,}35) \cdot (-0{,}25)}{1 - (-0{,}25)^2}} = \sqrt{0{,}497} = 0{,}705.$$

F-критерий Фишера:

$$F_{\text{факт}} = \frac{R^2}{1 - R^2} \cdot \frac{n - m - 1}{m} = \frac{0{,}497}{1 - 0{,}497} \cdot \frac{25 - 2 - 1}{2} = \frac{0{,}497}{0{,}503} \cdot 11 = 10{,}87.$$

Табличное значение критерия Фишера $F_{\text{таб}}(0{,}05; 23; 2) = 3{,}44$. $F_{\text{факт}} > F_{\text{таб}}$, следовательно, признается статистическая значимость уравнения в целом.

2) значимость коэффициентов множественной регрессии оценим с помощью *t*-критерия Стьюдента

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

$$F_{x_1} = \frac{R^2_{yx_1x_2} - r^2_{yx_1}}{1 - R^2_{yx_1x_2}} \cdot (n - m - 1) = \frac{0{,}497 - 0{,}4624}{1 - 0{,}497} \cdot 22 = 1{,}513.$$

$$F_{x_2} = \frac{R^2_{yx_1x_2} - r^2_{yx_2}}{1 - R^2_{yx_1x_2}} \cdot (n - m - 1) = \frac{0{,}497 - 0{,}1225}{1 - 0{,}497} \cdot 22 = 16{,}38$$

$$t_{b_1} = \sqrt{F_{x_1}} = \sqrt{1{,}513} = 1{,}23.$$

$$t_{b_2} = \sqrt{F_{x_2}} = \sqrt{16{,}38} = 4{,}05.$$

Табличное значение $t_{\text{таб}}(0{,}05; 25 - 2 - 1) = 3{,}1040$.

Так как $t_{b_1} < t_{\text{таб}}$, то коэффициент b_1 статистически не значим, а $t_{b_2} > t_{\text{таб}}$, значит коэффициент b_2 статистические значим.

Задача 4. В результате МНК-оценивания параметров модели на основе данных о 1000 работниках исследователь получил следующее уравнение:

$$\underset{}{\hat{y}} = \underset{(0{,}3)}{4{,}2} + \underset{(0{,}1)}{2{,}1} \cdot x - \underset{(0{,}2)}{3{,}5} \cdot d,$$

где *y* – зарплата работника в денежных единицах в час, x – стаж работы работника в годах, d – фиктивная переменная, которая равна единице, если работник – женщина, и равна нулю, если мужчина. В скобках приведены стандартные ошибки.

Дайте интерпретацию коэффициентам модели. Подтверждается ли на рассматриваемом рынке труда дискриминации по гендерному признаку?

Решение

Коэффициент перед переменной *x* означает, что при увеличении стажа работы на 1 год зарплата работника в среднем вырастет на 2,1 денежных единиц. Коэффициент перед переменной d означает, что если работник женщина ($d = 1$), то при одинаковом стаже она получает зарплату в среднем на 3,5 денежных единицы меньше.

Найдём значения *t*-статистик и сравним с табличным значением $t_{\text{таб}}(0{,}05; 1000 - 2 - 1) = 1{,}96$.

$$t_a = \frac{4{,}2}{0{,}3} = 14 > t_{\text{таб}}, \qquad t_{b_1} = \frac{2{,}1}{0{,}1} = 21 > t_{\text{таб}},$$

$$t_{b_1} = \frac{|-3{,}5|}{0{,}2} = 17{,}5 > t_{\text{таб}}.$$

Все параметры уравнения статистически значимы, в частности коэффициент перед фиктивной переменной, следовательно, дискриминации по гендерному признаку подтверждается.

Построение множественной регрессии с использованием табличного процессора Excel

1. Руководство авиакомпании по результатам анализа деятельности 15 своих представительств получило следующие данные за месяц:

Таблица 18

y	x_1	x_2	x_3
79,3	2,5	10	3
200,1	5,5	8	6
163,2	6	12	9
200,1	7,9	7	16
146	5,2	8	15
177,7	7,6	12	9
30,9	2	12	8
291,9	9	5	10
160	4	8	4
339,4	9,6	5	16
159,6	5,5	11	7
88,3	3	12	8
237,5	6	6	10
107,2	5	10	4
155	3,5	10	4

где y – общий доход от проданных билетов, млн. руб.;

x_1 – средства на развитие компании в регионе, млн. руб.;

x_2 – число конкурирующих компаний;

x_3 – процент пассажиров летавших бесплатно.

Задание:

1. Проверить наличие коллинеарности. Отобрать неколлинеарные факторы.

2. Построить уравнение линейной регрессии.

3. Определить коэффициент множественной корреляции, множественной детерминации, скорректированный коэффициент, вычислить частные коэффициенты корреляции.

4. С помощью F-критерия Фишера проверить значимость уравнения при уровне значимости 0,05.

5. С помощью t-критерия оценить статистическую значимость коэффициентов чистой регрессии.

6. Построить уравнение линейной регрессии в стандартизированном масштабе. Определить средние коэффициенты эластичности. Ранжировать факторы по степени их влияния на результат.

7. С помощью частного *F*-критерия Фишера проверить целесообразность включения факторов в модель.

8. Построить уравнение регрессии с учетом только информативных факторов.

9. Сделать выводы. Оформить в виде отчёта.

Решение

Так как оценка параметров уравнения множественной регрессии достаточно трудоёмкий процесс, то для проведения всех расчётов воспользуемся табличным процессором MS Excel.

Предварительно необходимо ввести исходные данные на лист Excel (4 столбца: y, x_1, x_2, x_3).

1. Построим корреляционную матрицу 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**.

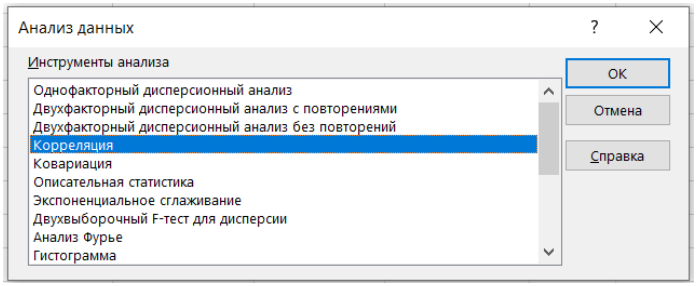


Рис. 19. Инструменты Анализа данных.

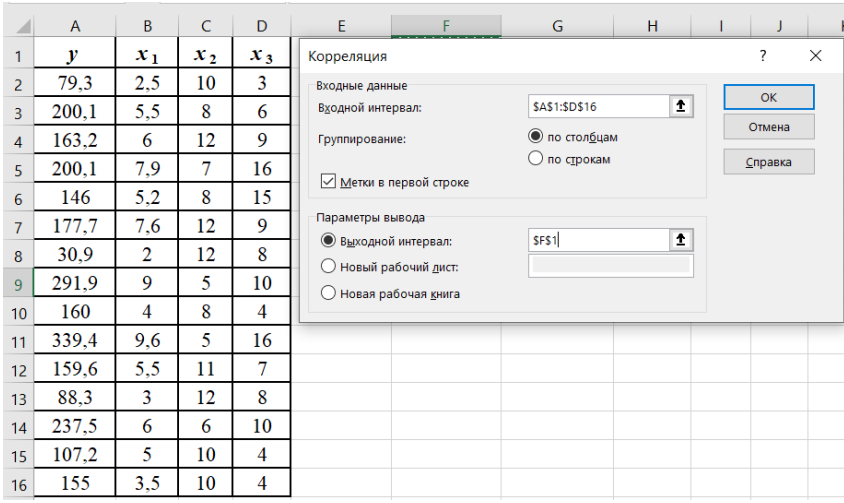
Задаём соответствующие диапазоны данных в появившемся диалоговом окне.

Входной интервал – диапазон ячеек со значениями всех переменных (выделяем вместе с заголовками столбцов).

Группирование – устанавливаем переключатель в положение «по столбцам» (анализируемые данные сгруппированы в столбцы).

Выходной интервал – ссылка на начальную ячейку диапазона, в который будет выведена матрицы. Размер диапазона определится автоматически.

Метки в первой строке – установим флажок (так как входной интервал содержит заголовки).



	A	B	C	D
1	y	x_1	x_2	x_3
2	79,3	2,5	10	3
3	200,1	5,5	8	6
4	163,2	6	12	9
5	200,1	7,9	7	16
6	146	5,2	8	15
7	177,7	7,6	12	9
8	30,9	2	12	8
9	291,9	9	5	10
10	160	4	8	4
11	339,4	9,6	5	16
12	159,6	5,5	11	7
13	88,3	3	12	8
14	237,5	6	6	10
15	107,2	5	10	4
16	155	3,5	10	4

Рис. 20. Диалоговое окно инструмента Корреляция.

Нажимаем ОК. Получим корреляционную матрицу:

	y	*x1*	*x2*	*x3*
y	1,00			
x1	0,89	1,00		
x2	-0,80	-0,62	1,00	
x3	0,53	0,66	-0,47	1,00

Рис. 21. Корреляционная матрица

Из матрицы следует, что $|r_{x_1x_2}| = |-0{,}62| < 0{,}7$; $r_{x_1x_3} = 0{,}66 < 0{,}7$; $|r_{x_2x_3}| = |-0{,}47| < 0{,}7$. Следовательно, коллинеарность между факторами отсутствует и нет основания исключать какой-либо фактор из рассмотрения.

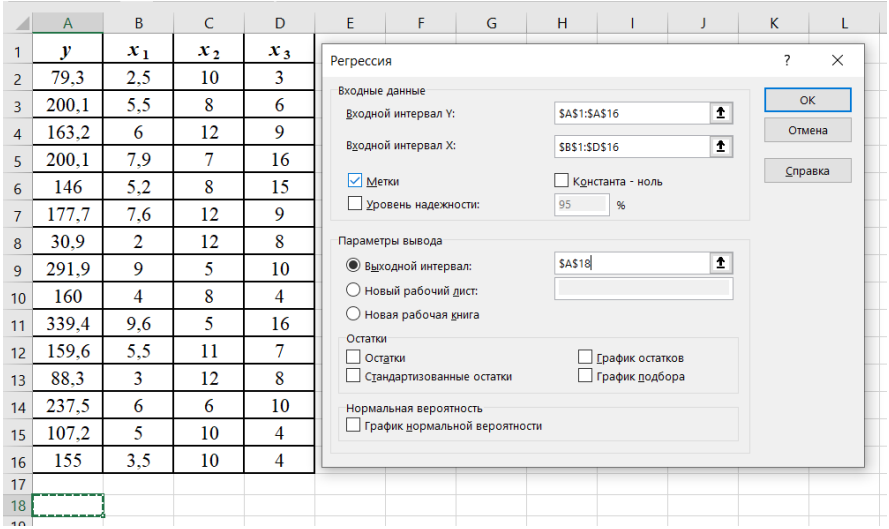
Парный коэффициент корреляции $r_{yx_1} = 0{,}89$ указывает на достаточно тесную связь общего дохода от проданных билетов со средствами на развитие компании в регионе x_1.

Парный коэффициент корреляции $r_{yx_2} = -0{,}80$ так же указывает на достаточно тесную связь общего дохода от проданных билетов с числом конкурирующих компаний x_2.

Парный коэффициент корреляции $r_{yx_3} = 0{,}53$ указывает на умеренную связь общего дохода от проданных билетов с процентом пассажиров летавших бесплатно x_3.

Все три фактора являются информативными. Таким образом, далее будет строиться регрессия y по факторам, x_1, x_2, x_3.

2. Для построения уравнения линейной регрессии используем инструмент «Данные. Анализ данных. Регрессия». Задав соответствующие диапазоны данных в окне,



	A	B	C	D
1	y	x_1	x_2	x_3
2	79,3	2,5	10	3
3	200,1	5,5	8	6
4	163,2	6	12	9
5	200,1	7,9	7	16
6	146	5,2	8	15
7	177,7	7,6	12	9
8	30,9	2	12	8
9	291,9	9	5	10
10	160	4	8	4
11	339,4	9,6	5	16
12	159,6	5,5	11	7
13	88,3	3	12	8
14	237,5	6	6	10
15	107,2	5	10	4
16	155	3,5	10	4
17				
18				

Рис. 22. Диалоговое окно инструмента Регрессия.

получим набор таблиц (рис. 23)

18	ВЫВОД ИТОГОВ						
19							
20	*Регрессионная статистика*						
21	Множественный R	0,951173773					
22	R-квадрат	0,904731546					
23	Нормированный R-квад	0,878749241					
24	Стандартная ошибка	27,79777068					
25	Наблюдения	15					
26							
27	Дисперсионный анализ						
28		*df*	*SS*	*MS*	*F*	*Значимость F*	
29	Регрессия	3	80720,3874	26906,7958	34,82106477	6,55432E-06	
30	Остаток	11	8499,876604	772,7160549			
31	Итого	14	89220,264				
32							
33		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
34	Y-пересечение	170,760025	52,08624542	3,278409177	0,007354521	56,11897175	285,4010782
35	x1	25,42327255	4,864201142	5,226607988	0,000282606	14,71723802	36,12930708
36	x2	-13,00348412	3,727756248	-3,488287124	0,005073845	-21,20822031	-4,79874794
37	x3	-2,705905925	2,315095982	-1,168809391	0,26718628	-7,801397826	2,389585976
38							

Рис. 23. Вывод итогов инструмента Регрессия.

Из последней таблицы следует, что уравнение регрессии имеет вид (значения из таблицы округлим до сотых):

$$\hat{y} = 170{,}76 + 25{,}42 \cdot x_1 - 13 \cdot x_2 - 2{,}71 \cdot x_3.$$

Параметр $b_1 = 25{,}42$ означает, что при увеличении средств на развитие компании в регионе на 1 млн. руб., при неизменном уровне других факторов, общий доход от проданных билетов увеличится в среднем на 25,42 млн. руб.

Параметр $b_2 = -13$ означает, что при увеличении числа конкурирующих компаний на единицу, при неизменном уровне других факторов, общий доход от проданных билетов уменьшится в среднем на 13 млн. руб.

Параметр $b_3 = -2{,}71$ означает, что при увеличении числа пассажиров летавших бесплатно на 1%, при неизменном уровне других факторов, общий доход от проданных билетов уменьшится в среднем на 2,71 млн. руб.

3. Коэффициенты множественной корреляции и детерминации определяются из первой таблицы. $R_{yx_1x_2x_3}$= 0,95 означает, что между общим доходом от проданных билетов и полным набором факторов связь очень тесная. $R^2_{yx_1x_2x_3}$= 0,9 означает, что вариация общего дохода на 90% объясняется вариацией включённых в модель факторов. Скорректированный коэффициент детерминации $\hat{R}^2 = 0{,}88$ практически равен $R^2_{yx_1x_2x_3}$, значит теснота связи не является преувеличенной.

Чтобы вычислить частные коэффициенты корреляции предварительно необходимо найти $R^2_{yx_2x_3}$, $R^2_{yx_1x_3}$, $R^2_{yx_1x_2}$. Для этого можно воспользоваться инструментом «Регрессия» или функцией ЛИНЕЙН.

Например, для определения $R^2_{yx_2x_3}$ скопируем на новый лист столбцы y, x_2, x_3 и запустим инструмент Регрессия.

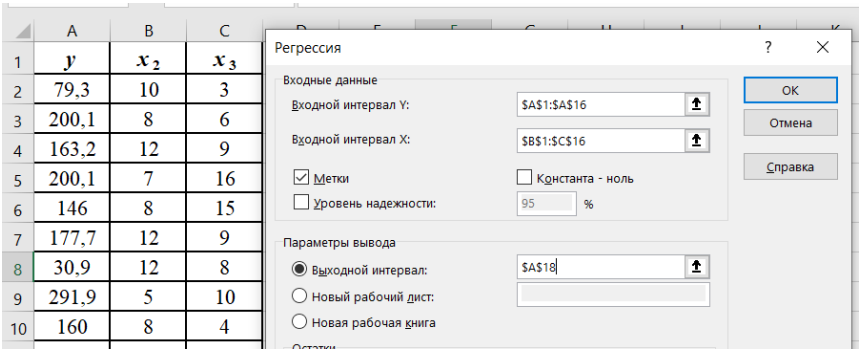


Рис. 24. Диалоговое окно инструмента Регрессия для модели с двумя факторами x_2, x_3.

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,817399
R-квадрат	0,668142
Нормированный R-квадрат	0,612832
Стандартная ошибка	49,67267
Наблюдения	15

Рис. 25. Фрагмент вывода итогов инструмента Регрессия.

Получим значение $R^2_{yx_2x_3} = 0{,}67$.

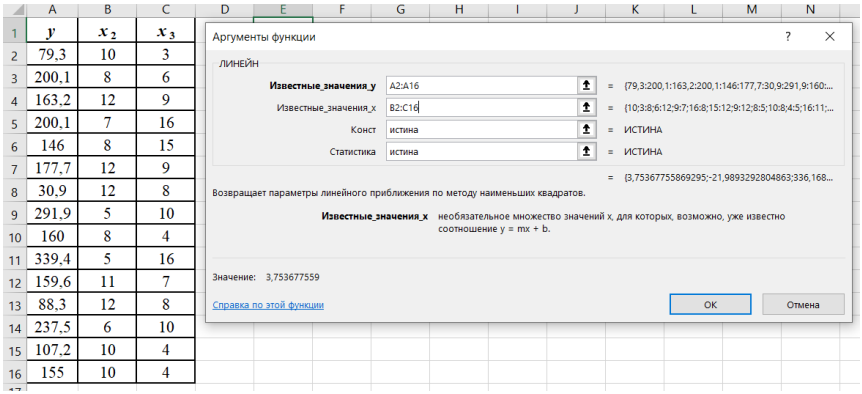
Или воспользуемся функцией ЛИНЕЙН.

В случае множественной регрессии $\hat{y} = a + b_1 \cdot x_1 + b_2 \cdot x_2$, когда значения y зависят от двух переменных, функция ЛИНЕЙН возвращает 12 статистик в следующем формате (табл. 19):

Таблица 19

Значение коэффициента b_2	Значение коэффициента b_1	Значение свободного члена a
Стандартная ошибка b_2 m_{b_2}	Стандартная ошибка b_1 m_{b_1}	Стандартная ошибка a m_a
Коэффициент детерминации R^2	Оценка стандартного отклонения остатков $S_{\text{ост}}$	Н/Д
Значение F-статистики	Число степеней свободы	Н/Д
Регрессионная сумма квадратов $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	Остаточная сумма квадратов $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	Н/Д

Выделяем диапазон размером 5 × 3 (5 строк, 3 столбца). Набираем формулу {=ЛИНЕЙН(A2:A16;B2:C16;ИСТИНА;ИСТИНА)}. Либо заполняем аргументы функции в диалоговом окне



	A	B	C
1	y	x_2	x_3
2	79,3	10	3
3	200,1	8	6
4	163,2	12	9
5	200,1	7	16
6	146	8	15
7	177,7	12	9
8	30,9	12	8
9	291,9	5	10
10	160	8	4
11	339,4	5	16
12	159,6	11	7
13	88,3	12	8
14	237,5	6	10
15	107,2	10	4
16	155	10	4

Рис. 26. Диалоговое окно функции ЛИНЕЙН.

Для аргумента **Известные_значения_x**, выделяем оба столбца значений x из диапазона B2:C16. *Диапазоны выделяем без заголовков*. Нажимаем ОК.

Далее нажимаем клавишу F2, а затем комбинацию клавиш Ctrl+Shift+Enter. В результате получим

3,753678	-21,9893	336,1683
3,498101	5,910485	73,92412
0,668142	49,67267	#Н/Д
12,08	12	#Н/Д
59611,77	29608,49	#Н/Д

Рис. 27. Результат функции ЛИНЕЙН.

Откуда находим значение $R^2_{yx_2x_3} = 0{,}67$.

Аналогично получаем два других коэффициента детерминации: $R^2_{yx_1x_3} = 0{,}8$; $R^2_{yx_1x_2} = 0{,}89$.

Частные коэффициенты корреляции:

$$r_{yx_1 \cdot x_2x_3} = \sqrt{1 - \frac{1 - R^2_{yx_1x_2x_3}}{1 - R^2_{yx_2x_3}}} = \sqrt{1 - \frac{1 - 0{,}9}{1 - 0{,}67}} = 0{,}84;$$

$$r_{yx_2 \cdot x_1x_3} = \sqrt{1 - \frac{1 - R^2_{yx_1x_2x_3}}{1 - R^2_{yx_1x_3}}} = \sqrt{1 - \frac{1 - 0{,}9}{1 - 0{,}8}} = 0{,}72;$$

$$r_{yx_3 \cdot x_1x_2} = \sqrt{1 - \frac{1 - R^2_{yx_1x_2x_3}}{1 - R^2_{yx_1x_2}}} = \sqrt{1 - \frac{1 - 0{,}9}{1 - 0{,}89}} = 0{,}33.$$

Сравнивая частные коэффициенты корреляции с парными можно сделать вывод, что парный коэффициент корреляции r_{yx_3} даёт завышенную ($r_{yx_3} = 0{,}53 > 0{,}33$), а r_{yx_1} и r_{yx_2} немного завышенную оценку тесноты связи ($r_{yx_1} = 0{,}89 > 0{,}84$, $|r_{yx_2}| = 0{,}79 > 0{,}72$).

4. Проверка значимости уравнения регрессии основана на использовании *F*-критерия Фишера. Фактическое значение критерия берётся из второй таблицы, т. е.

$$F_{\text{факт}} = 32{,}82.$$

Для определения табличного значения используем встроенную функцию MS Excel F.ОБР.ПХ (или FРАСПОБР). В свободной ячейке набираем формулу «=F.ОБР.ПХ(0,05;3;11)» либо заполняем аргументы функции в диалоговом окне: **Вероятность** = 0,05, **Степени_свободы1** = 3, **Степени_свободы2** = 15 – 3 – 1 = 11. В результате получаем $F_{\text{табл}} = 3{,}5874$.

Так как $F_{\text{факт}} > F_{\text{табл}}$, то уравнение регрессии статистически значимо. В ячейке «*Значимость F*» число 6,554322E-06 = 0,000006554322 меньше

0,05 из чего также следует, что уравнение регрессии статистически значимо.

5. Фактические значения *t*-критерия берём из третьей таблицы (значения берутся по модулю). $t_{b_1} = 5{,}23$; $t_{b_2} = 3{,}49$; $t_{b_3} = 1{,}17$. Табличное значение найдём при помощи встроенной функции СТЬЮДЕНТ.ОБР.2Х или для более ранних версий Excel функцию СТЬЮДРАСПОБР. Набираем формулу «=СТЬЮДЕНТ.ОБР.2Х(0,05;11)» либо заполняем аргументы функции в диалоговом окне (рис. 28). В результате получаем $t_{\text{табл}} = 2{,}2$.

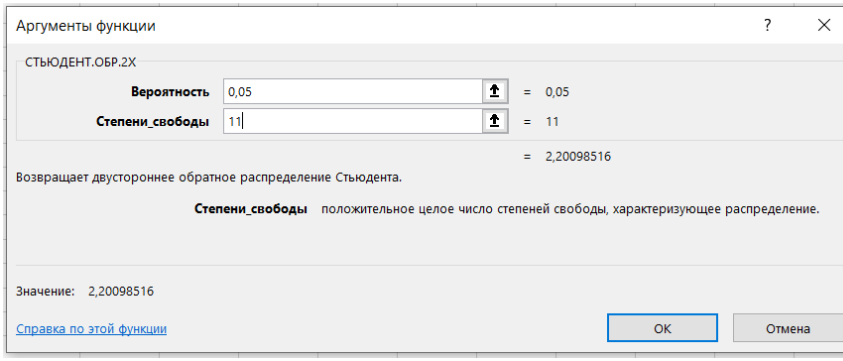


Рис. 28. Диалоговое окно функции СТЬЮДЕНТ.ОБР.2Х.

$t_{b_1} > t_{\text{таб}}$ – параметр b_1 значим;

$t_{b_2} > t_{\text{таб}}$– параметр b_2 значим;

$t_{b_3} < t_{\text{таб}}$ – параметр b_3 не значим.

В столбце «*P-Значение*» числа в строках x_1 (0,00028) и x_2 (0,005) меньше 0,05 а в строке x_3 (0,267) больше 0,05 из чего также следует, что – параметры b_1 и b_2 значимы, а параметр b_3 не значим.

6. Стандартизированные коэффициенты связаны с коэффициентами чистой регрессии следующими соотношениями:

$$\beta_j = b_j \frac{\sigma_{x_j}}{\sigma_y}.$$

Средние квадратические отклонения σ_y, σ_{x_i} определим, используя функцию MS Excel «СТАНДОТКЛОН»: $\sigma_y = 79{,}83$; $\sigma_{x_1} = 2{,}3$; $\sigma_{x_2} = 2{,}55$; $\sigma_{x_3} = 4{,}31$.

$$\beta_1 = 25{,}42\frac{2{,}3}{79{,}83} = 0{,}73;\ \beta_2 = -13\frac{2{,}55}{79{,}83} = -0{,}42;\ \beta_3 = -2{,}7\frac{4{,}31}{79{,}83} = -0{,}15.$$

Уравнение регрессии в стандартизированном масштабе имеет вид

$$t_y = 0{,}73t_{x1} - 0{,}42t_{x2} - 0{,}15t_{x3}.$$

Средние коэффициенты эластичности $\bar{Э}_j = b_j \frac{\bar{x}_j}{\bar{y}}$.

Средние значения переменных найдём с помощью функции «СРЗНАЧ».

$$\bar{Э}_1 = 25{,}42\frac{5{,}49}{169{,}08} = 0{,}82; \qquad \bar{Э}_2 = -13\frac{9{,}07}{169{,}08} = -0{,}7;$$

$$\bar{Э}_3 = -2{,}7\frac{8{,}6}{169{,}08} = -0{,}14.$$

Так как $\beta_1 > |\beta_2| > |\beta_3|$, то самое сильное влияние на доход от продажи билетов оказывает объём средств, направленных на развитие, самое слабое – процент пассажиров, летавших бесплатно. Средние коэффициенты эластичности это подтверждают.

7. Частные критерии Фишера:

$$F_{x_1} = \frac{R^2_{yx_1x_2x_3} - R^2_{yx_2x_3}}{1 - R^2_{yx_1x_2x_3}} \cdot \frac{n-m-1}{1} = \frac{0{,}9 - 0{,}668}{1 - 0{,}9} \cdot \frac{15-3-1}{1} = 27{,}32$$

или $F_{x_1} = (t_{b_1})^2 = 27{,}32$;

$$F_{x_2} = \frac{R^2_{yx_1x_2x_3} - R^2_{yx_1x_3}}{1 - R^2_{yx_1x_2x_3}} \cdot \frac{n-m-1}{1} = \frac{0{,}9 - 0{,}799}{1 - 0{,}9} \cdot \frac{15-3-1}{1} = 12{,}17$$

или $F_{x_2} = (t_{b_2})^2 = 12{,}17$;

$$F_{x_3} = \frac{R^2_{yx_1x_2x_3} - R^2_{yx_1x_2}}{1 - R^2_{yx_1x_2x_3}} \cdot \frac{n-m-1}{1} = \frac{0{,}9 - 0{,}89}{1 - 0{,}9} \cdot \frac{15-3-1}{1} = 1{,}37$$

или $F_{x_3} = (t_{b_3})^2 = 1{,}37$;

$F_{\text{таб}}(0{,}05;\ 1;\ 11) = 4{,}84$.

Сравнивая фактические значения с табличным, делаем вывод, что включение в модель фактора x_3 не целесообразно.

Замечание. Расчёты желательно выполнить в Excel. Множественные коэффициенты детерминации могут быть очень близки по значению и при ручном счёте из-за округлений можно получить нулевые значения.

8. Фактор x_3 оказывает на y самое слабое влияние; параметр b_3 статистически не значим и по критерию Фишера включение этого фактора в модель не целесообразно. Поэтому фактор x_3 из модели нужно исключить. Уравнение регрессии нужно пересчитать, используя только два фактора x_1 и x_2. Сделаем это снова с помощью инструмента **Регрессия**. Получим

ВЫВОД ИТОГОВ						
Регрессионная статистика						
Множественный R	0,9449					
R-квадрат	0,8929					
Нормированный R-квадрат	0,8750					
Стандартная ошибка	28,2186					
Наблюдения	15,0000					
Дисперсионный анализ						
	df	*SS*	*MS*	*F*	*Значимость F*	
Регрессия	2	79664,7681	39832,38405	50,02237598	1,50917E-06	
Остаток	12	9555,4959	796,291325			
Итого	14	89220,264				
	Коэффициент	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	159,862921	52,02089806	3,073051921	0,00966185	46,51912063	273,2067209
x1	22,3881879	4,175353964	5,361985624	0,00017026	13,29087314	31,48550271
x2	-12,531556	3,761931808	-3,331149285	0,005985353	-20,72810174	-4,335011164

Рис.29. Вывод итогов для модели с двумя значащими факторами.

Уравнение регрессии имеет вид

$$\hat{y} = 159{,}86 + 22{,}39 \cdot x_1 - 12{,}53 \cdot x_2.$$

$R_{yx_1x_2} = 0{,}94$ коэффициент корреляции незначительно сократился.

$R^2_{yx_1x_2} = 0{,}89$. После исключения фактора x_3 (процент пассажиров летавших бесплатно) доля объяснённой дисперсии сократилась всего на 1%.

На основании F-критерия и t-статистики делаем вывод, что и уравнение в целом и все его параметры статистически значимы.

Задачи для самостоятельного решения

1. При оценивании модели пространственной выборки с помощью МНК по 20 наблюдениям получено следующее уравнение

$$\hat{y} = \underset{(0,4)}{12} + \underset{(0,35)}{4,23x_1} - \underset{(0,1)}{0,43x_2}, \qquad R^2 = 0,8,$$

В скобках указаны стандартные ошибки.

Оцените значимость коэффициентов модели, значимость уравнения в целом.

2. По заданному уравнению регрессии

$$\hat{y} = 25 + 2 \cdot x_1 + 5,5 \cdot x_2$$

найти средние коэффициенты эластичности, если $\bar{x}_1 = 4$, $\bar{x}_2 = 6$, $\bar{y} = 10$.

3. По ежемесячным данным за 3 года построена выборочная множественная регрессия $\hat{y} = -12,23 + 0,91x_1 - 2,1x_2$, где y – объем потребления (в сотнях руб.), x_1 – располагаемый доход (в тыс. руб.), x_2 – процент банковской ставки по вкладам. Известны значения t-статистик коэффициентов: $t_{b_1} = 12,7$, $t_{b_2} = -3,2$ и коэффициент детерминации $R^2 = 0,96$.

Дайте интерпретацию коэффициентам. Проверьте статистическую значимость коэффициентов при $\alpha = 0,05$. Найдите стандартные ошибки коэффициентов. Проверьте статистическую значимость коэффициента детерминации (значимость уравнения в целом по критерию Фишера).

4. Зависимость среднего душевого дохода (в руб.) от средней заработной платы в день (в руб.) и среднего возраста (в годах) отражена уравнением множественной регрессии $\hat{y} = -73,52 + 1,62x_1 - 2,25x_2$. Число наблюдений $n = 30$. Известны выборочные коэффициенты корреляции $r_{yx_1} = 0,84$, $r_{yx_2} = -0,21$, $r_{x_1x_2} = -0,116$.

Задание:

1) Дайте интерпретацию коэффициентам уравнения.

2) Найдите уравнение множественной регрессии в стандартизированном масштабе.

3) Вычислите частные коэффициенты корреляции, сравните их с парными.

4) Рассчитайте множественный коэффициент корреляции.

5) Рассчитать общий и частные критерии Фишера, сделать выводы.

5. Задано уравнение регрессии в стандартизованных переменных

$$\hat{t}_y = -0,82t_{x_1} + 0,65t_{x_2} - 0,43t_{x_3}.$$

При этом вариации всех переменных равны следующим величинам: $V_y = 32\%;\ V_{x_1} = 38\%;\ V_{x_2} = 43\%;\ V_{x_3} = 35\%.$

Задание: сравнить факторы по степени влияния на результирующий признак и определить значения средних коэффициентов эластичности.

6. Перейти от уравнения регрессии в натуральном масштабе переменных, описывающей зависимость среднедневного душевого дохода (y, руб.) от среднедневной заработной платы одного работающего (x_1, руб.) и среднего возраста безработного (x_2, лет)

$$\hat{y} = 337{,}37 + 1{,}976x_1 - 12{,}09x_2,$$

к уравнению регрессии в стандартизованном масштабе переменных, если известно, что $\sigma_y = 61{,}44,\ \sigma_{x_1} = 25{,}86,\ \sigma_{x_2} = 0{,}58$ и интерпретировать коэффициенты уравнения регрессии.

7. По 30 наблюдениям получены следующие данные:

уравнение регрессии

$$\hat{y} = a + 0{,}18 \cdot x_1 + 0{,}015 \cdot x_2 - 7{,}5 \cdot x_3;$$

средние значения

$$\bar{y} = 200, \bar{x}_1 = 150, \bar{x}_2 = 20, \bar{x}_3 = 100;$$

средние квадратические отклонения

$$\sigma_y = 18, \sigma_{x_1} = 10, \sigma_{x_2} = 1{,}7, \sigma_{x_3} = 2{,}1.$$

Задание:

1) Оцените параметр a.

2) Ранжировать факторы по силе влияния на результат с помощью стандартизированных коэффициентов.

8. Было обследовано 20 предприятий по 3 показателям. (в скобках указаны стандартные ошибки)

$$\hat{y} = \underset{(7{,}3)}{-10{,}5} + \underset{(3{,}2)}{6{,}57x_1} - \underset{(1{,}3)}{0{,}22x_2} + \underset{(3{,}3)}{7{,}8x_3}$$

Оцените целесообразность включения факторов в модель с помощью частного F-критерия Фишера.

9. По 30 наблюдениям получено уравнение множественной регрессии $\hat{y} = -73{,}52 + 1{,}62x_1 - 2{,}25x_2$, y – средний душевой доход (в руб.) x_1 – средняя заработная плата в день (в руб.) x_2 – средний возраст (лет); вычислены выборочные парные коэффициенты корреляции $r_{yx_1} = 0{,}84$, $r_{yx_2} = -0{,}21$, $r_{x_1x_2} = -0{,}116$ и коэффициент детерминации $R^2 = 0{,}72$.

Проверьте статистическую значимость уравнения в целом по критерию Фишера. По частному F-критерию проверьте целесообразность

включения в уравнение фактора x_2 после фактора x_1 и наоборот – фактора x_1 после фактора x_2.

10. По 25 предприятиям легкой промышленности получена информация, характеризующая зависимость объема выпуска продукции y (млн. руб.) от количества отработанных за год человеком часов x_1 (тыс. чел./час) и среднегодовой стоимости производственного оборудования x_2 (млн. руб.):

Уравнение регрессии $\hat{y} = 30 + 0{,}05x_1 + 2{,}56x_2$.

Множественный коэффициент корреляции 0,85.

Сумма квадратов отклонений фактических значений результативного признака от расчетных 210.

Задание:

1) Определите коэффициент детерминации в этой модели.

2) Составьте таблицу дисперсионного анализа.

11. Имеется информация по 30 наблюдениям

Признак	Среднее значение	Коэффициент вариации, %	Уравнение регрессии
y	30	22	$\hat{y} = 15 + 2x_1 - 3x_2$
x_1	18	25	$\hat{y} = 7 + 1{,}3x_1$
x_2	10	12	$\hat{y} = 5 - 3{,}1x_2$

Оценить значимость 1-го уравнения в целом и значимость его коэффициентов, если известно, что $r_{x1x2} = -0{,}36$.

12. Изучается зависимость прибыли y (тыс. руб.) от выработки продукции на одного работника x_1 (ед.) и индекса цен на продукцию x_2 (%). Средние значения: $\bar{y} = 250, \bar{x}_1 = 45, \bar{x}_2 = 120$.

Средние квадратические отклонения $\sigma_y = 15, \sigma_{x_1} = 11, \sigma_{x_2} = 14$.

Парные коэффициенты корреляции $r_{yx_1} = 0{,}7,\ r_{yx_2} = 0{,}65;\ r_{x_1x_2} = 0{,}21$.

Число наблюдений $n = 30$.

Задание:

1) Найти уравнение множественной регрессии в стандартизированном масштабе. Дать интерпретацию коэффициентам.

2) Найти линейное уравнение множественной регрессии. Дать интерпретацию коэффициентам.

3) Рассчитать множественный и частные коэффициенты корреляции.

4) Рассчитать общий и частные критерии Фишера, сделать выводы.

13. Коэффициенты парной корреляции равны: $r_{yx_1} = 0{,}8$, $r_{yx_2} = -0{,}7$, $r_{x_1x_2} = 0{,}2$. Чему равен коэффициент множественной корреляции $R_{yx_1x_2}$?

14. Заданы следующие коэффициенты парной корреляции между факторами x_1, x_2, x_3, x_4:

$$r_{12} = 0{,}42;\ r_{14} = 0{,}25;\ r_{23} = 0{,}3;\ r_{31} = 0{,}51; r_{34} = 0{,}52; r_{42} = 0{,}28.$$

Построить полную матрицу коэффициентов корреляции всех пар переменных. Проверить наличие мультиколлинеарности.

15. В таблице указаны парные коэффициенты корреляции. Проведите анализ целесообразности включения заданных факторов в уравнение множественной линейной регрессии.

	y	x_1	x_2	x_3	x_4
y	1				
x_1	0,81	1			
x_2	0,48	0,18	1		
x_3	0,13	0,73	0,15	1	
x_4	0,02	0,15	0,5	0,25	1

16. Результаты расчетов параметров выборочной множественной регрессии с помощью функции ЛИНЕЙН приведены в таблице 20:

Таблица 20

26,87	17,33	88,65
8,07	2,82	29,90
0,84	49,65	#Н/Д
70,82	27	#Н/Д
349186	66568	#Н/Д

Задание:

1) Запишите уравнение регрессии.

2) Проверьте значимость параметров выборочной регрессии и всего уравнения в целом.

3) Постройте доверительные 95 % интервалы параметров.

4) Постройте прогноз для значений объясняющих переменных $x_1 = 10{,}83,\ x_2 = 4{,}73$.

17. Зависимость урожайности пшеницы y (ц/га) от количества внесенных минеральных удобрений на 1га x_1 (ц) и осадков x_2 (мм) характеризуется следующим уравнением:

$$\hat{y} = -110 + 0{,}15x_1 - 0{,}007x_1^2 + 0{,}7x_2 - 0{,}001x_2^2.$$

При этом $\sigma_y = 2{,}1; n = 25; R = 0{,}9$.

Составить таблицу дисперсионного анализа. Проверить статистическую значимость уравнения множественной регрессии при уровне значимости $\alpha = 0{,}05$.

18. Информация о деятельности 15 магазинов, входящих в сеть торгового предприятия, представлена в таблице 21.

Таблица 21

№ магазина	Годовой товарооборот, млн. руб.	Торговая площадь, тыс. м^2	Среднее число посетителей в день, тыс. чел.
1	19,75	0,23	8,25
2	36,09	0,31	9,44
3	40,95	0,55	9,31
4	41,08	0,48	11,01
5	56,29	0,78	8,54
6	68,51	0,98	7,51
7	75,01	0,94	12,36
8	89,05	1,21	10,81
9	91,13	1,29	9,89
10	91,26	1,12	13,72
11	99,84	1,29	12,27
12	108,55	1,49	13,92
13	110,25	1,45	14,02
14	112,32	1,52	14,13
15	115,68	1,61	14,84

Оцените с помощью метода наименьших квадратов параметры линейного уравнения множественной регрессии. Дайте интерпретацию полученным оценкам.

19. Зависимость объёма выпуска продукции y (млн. руб.) от численности занятых на предприятии x_1 (чел.) и среднегодовой стоимости основных фондов x_2 (млн. руб.)

$$\ln y = 4 + 0{,}52 \cdot \ln x_1 + 0{,}66 \cdot \ln x_2 .$$

Задание:

1) Представить данное уравнение в естественной форме (не в логарифмах).

2) Интерпретировать параметры b_1 и b_2.

3) Оценить значимость параметров данного уравнения, если известны значения *t*-критерия: $t_{b_1} = 2{,}13$, $t_{b_2} = 3{,}41$ и $n = 20$.

20. По 45 наблюдениям методом наименьших квадратов экономист оценил регрессию

$$\hat{y} = \underset{(3,5)}{-15} + \underset{(1,1)}{2,5} \cdot x_1 + \underset{(2,1)}{8} \cdot x_2,$$

где y – почасовая зарплата (ден. ед.),

x_1 – уровень образования (годы обучения),

x_2 – фиктивная переменная, равная 1, если индивидуум мужчина, и равная 0, если индивидуум женщина.

В скобках приведены стандартные ошибки.

Затем решил включить в число объясняющих переменных результаты ЕГЭ (из 100 баллов) – x_3. Получилась регрессия:

$$\hat{y} = \underset{(4,3)}{-18} + \underset{(1,2)}{2,7} \cdot x_1 + \underset{(1,8)}{6,1} \cdot x_2 + \underset{(0,2)}{0,07} \cdot x_3,$$

1) Влияет ли пол на зарплату? Если да, то как?

2) Влияет ли результат ЕГЭ на зарплату? Если да, то как?

21. По 25 наблюдениям оценено уравнение регрессии количества проданных в некоторой торговой точке бутылок газированной воды q от цены газировки p, средней температуры дня t. В модель также была включена дамми-переменная d, равная 1 для наблюдений, соответствующих выходным дням, и 0 для будних дней. Оценивание дало следующие результаты:

$$\hat{q} = \underset{(5,7)}{37,5} - \underset{(1,0)}{2,4} \cdot p + \underset{(1,5)}{2,1} \cdot t + \underset{(3,2)}{13,7} \cdot d, R^2 = 0,7$$

(в скобках под оценками коэффициентов указаны их стандартные ошибки).

а) Проверьте значимость регрессии на уровне значимости 5%.

б) Определите, какие из включённых в модель факторов оказывают значимое влияние на объём продаж? Используйте уровень значимости 1%.

в) Отличаются ли продажи в выходные и будни? Если да, то как?

Задачи 22 и 23 решить с помощью MS Excel.

22. По данным, представленным в таблице 22, исследуется зависимость между величиной накладных расходов 30 строительных организаций y (млн. руб.) и следующими тремя основными факторами:

x_1 – объемом выполненных работ, млн. руб.

x_2 – численностью рабочих, чел.

x_3 – фондом зарплаты, млн. руб.

Таблица 22

№	Накладные расходы, млн. руб.	Объем работ, млн. руб.	Численность рабочих, чел.	Фонд заработной платы рабочих, млн. руб.
1	3,40	15,0	670	5,76
2	4,10	14,7	622	6,10
3	4,10	13,3	566	6,06
4	3,10	14,6	518	4,92
5	2,80	11,7	510	4,13
6	2,10	10,6	452	4,38
7	2,50	10,0	447	4,16
8	2,00	9,0	497	4,32
9	2,40	9,5	428	4,02
10	2,30	7,0	381	3,32
11	2,40	9,1	385	3,62
12	2,50	6,8	412	3,46
13	2,20	5,5	293	2,14
14	1,60	5,1	284	2,24
15	3,40	12,2	514	3,96
16	2,70	11,0	407	3,34
17	3,20	9,3	577	3,68
18	2,90	5,9	265	2,12
19	4,80	25,9	977	10,65
20	3,70	23,5	724	6,81
21	4,40	19,8	983	9,24
22	3,70	18,8	828	8,86
23	4,80	19,1	766	7,35
24	3,70	18,8	615	5,29
25	3,60	17,4	583	5,83
26	4,00	14,1	591	6,27
27	3,80	13,80	593	5,40
28	4,2	14,8	579	6,33
29	4,6	15,1	623	6,92
30	3,4	17,6	596	5,32

Задание:

1) Построить линейную модель множественной регрессии, найти индекс корреляции, индекс детерминации; оценить значимость модели в целом (*F*-критерий Фишера) и отдельных ее параметров (*t*-критерий Стьюдента).
2) Проанализировать матрицу парных коэффициентов корреляции на наличие межфакторной корреляции и мультиколлинеарности. Отобрать неколлинеарные факторы.
3) Построить линейную модель регрессии только со значимыми факторами (на основании выводов, сделанных в п.п. 1 и 2). Дать экономическую интерпретацию коэффициентов модели. Оценить качество построенной модели (индексы корреляции и детерминации, *F*-критерий Фишера, средняя ошибка аппроксимации).

23. В таблице 22 представлены данные по Забайкальскому краю за 2001-2021 годы

Таблица 23

Год	Продукция сельского хозяйства (животноводства), млн. руб.	Крупный рогатый скот (тыс. голов)	Свиньи (тыс. голов)	Овцы и козы (тыс. голов)	Лошади (тыс. голов)	Птица (тыс. голов)
2001	4165,6	423,3	92,2	441,6	56,7	1077,1
2002	4872	435,6	114,2	481,7	57,7	1105,1
2003	5066,7	424,1	126,3	526,9	59,1	976
2004	5220	416,3	109	551	60,1	879,2
2005	5948,8	404,9	89,8	560,2	61,4	834,3
2006	7183,8	417	110,9	567,5	64,4	892,2
2007	8125,1	435,1	123,7	584,3	67,9	1201
2008	9017,9	442,4	114,6	593,7	73	804,9
2009	9765,6	439,2	109,9	553,8	74	715,9
2010	10724,6	441,5	126,1	520	74,2	767,7
2011	11847,9	457	131,5	522	76,3	753,7
2012	12780,3	474,2	122,3	520,7	79,5	711,2
2013	13499,8	471,8	115,2	494,7	83,1	716,2
2014	14182,7	479,7	115,4	484,6	78,4	704,2
2015	15487,4	469,5	113,4	468,4	76	665
2016	16441,9	453,1	104,1	455,6	72	667
2017	16465,6	451,4	91	466,7	73,1	656,1
2018	16357,8	452,8	68,8	496,2	98,2	451,4
2019	16683,9	454	71,4	468,6	98,9	459,6
2020	16919,8	455,9	65,8	445,7	102,7	428,5
2021	18135,2	457,7	51,2	415,5	107,6	406,8

Задание:

1) Проверить наличие коллинеарности. Отобрать неколлинеарные факторы.
2) Построить уравнение линейной регрессии.
3) Определить коэффициент множественной корреляции, множественной детерминации, скорректированный коэффициент.
4) С помощью *F*-критерия Фишера проверить значимость уравнения при уровне значимости 0,05.
5) С помощью *t*-критерия оценить статистическую значимость коэффициентов чистой регрессии.
6) Построить уравнение регрессии с учетом только информативных факторов.

Домашнее задание к разделу множественная регрессия

Исследуется взаимосвязь показателей качества жизни населения по выборке для 25 регионов:

y – средняя ожидаемая продолжительность жизни при рождении, лет;

x_1 – уровень рождаемости, чел. на 1000 чел. населения;

x_2 – доля населения с денежными доходами ниже величины прожиточного минимума, % от всего населения;

x_3 – объем социальных выплат, млрд. у. е.

Таблица 24

№	y	x_1	x_2	x_3	№	y	x_1	x_2	x_3
1	68,1	10,2+0,2*N*	11,2+0,1*N*	6,09	**14**	68,7	12,5+0,3*N*	13,0+0,2*N*	5,58
2	68,2	10,5+0,2*N*	14,0+0,1*N*	6,79	**15**	68,6	11,2+0,3*N*	15,1+0,2*N*	6,52
3	69,0	11,7+0,2*N*	11,9+0,1*N*	4,50	**16**	68,6	12,5+0,3*N*	12,8+0,2*N*	5,70
4	68,2	11,3+0,2*N*	12,0+0,1*N*	4,71	**17**	69,0	12,2+0,3*N*	12,2+0,2*N*	5,72
5	66,6	8,8+0,2*N*	14,3+0,1*N*	5,72	**18**	68,5	10,5+0,3*N*	13,9+0,2*N*	6,84
6	68,6	11,9+0,2*N*	11,0+0,1*N*	4,69	**19**	67,9	10,9+0,3*N*	12,9+0,2*N*	5,43
7	68,3	11,4+0,2*N*	11,3+0,1*N*	6,11	**20**	69,7	13,1+0,3*N*	11,8+0,2*N*	6,02
8	67,3	9,0+0,2*N*	14,3+0,1*N*	6,65	**21**	68,5	10,4+0,3*N*	11,6+0,2*N*	5,11
9	68,6	11,4+0,2*N*	12,6+0,1*N*	5,18	**22**	68,6	11,9+0,3*N*	13,1+0,2*N*	5,34
10	68,4	12,0+0,2*N*	12,5+0,1*N*	5,41	**23**	68,3	12,5+0,3*N*	12,1+0,2*N*	4,95
11	69,1	11,1+0,2*N*	10,5+0,1*N*	5,83	**24**	67,0	8,1+0,3*N*	15,2+0,2*N*	7,43
12	69,1	12,3+0,2*N*	11,2+0,1*N*	4,85	**25**	68,0	10,1+0,3*N*	12,3+0,2*N*	6,06
13	68,8	12,0+0,2*N*	12,5+0,1*N*	5,57					

N – номер варианта.

Задание: На основании данных таблицы

1. Построить уравнение линейной регрессии.
2. Определить коэффициент множественной корреляции, множественной детерминации, скорректированный коэффициент, вычислить частные коэффициенты корреляции.
3. С помощью *F*-критерия Фишера проверить значимость уравнения при уровне значимости 0,05.
4. С помощью *t*-критерия оценить статистическую значимость коэффициентов чистой регрессии.
5. Построить уравнение линейной регрессии в стандартизированном масштабе. Определить средние частные коэффициенты эластичности. Ранжировать факторы по степени их влияния на результат.

6. С помощью частного *F*-критерия Фишера проверить целесообразность включения факторов в модель.

7. Построить уравнение регрессии с учетом только информативных факторов.

Указания к решению. При выполнении задания использовать возможности надстройки «Анализ данных» табличного процессора MS Excel.

Рекомендуемая литература

1. Доугерти, К. Введение в эконометрику / К. Доугерти. – М.: ИНФРА-М, 1997. – 402 с.
2. Колпаков, В. Ф. Экономико-математическое и эконометрическое моделирование: компьютерный практикум: учеб. пособие / В. Ф. Колпаков. – Москва: ИНФРА-М, 2018 – 396 с.
3. 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.
4. Методы эконометрики: Учебник / С.А. Айвазян; Московская школа экономики МГУ им. М.В. Ломоносова (МШЭ). – М.: Магистр: ИНФРА-М, 2010-512 с.
5. Практикум по эконометрике: учебное пособие / под ред. И. И. Елисеевой. – М.: Финансы и статистика, 2002. – 191 с.
6. Шанченко, Н. И. Лекции по эконометрике: учебное пособие для студентов высших учебных заведений, обучающихся по специальности «Прикладная информатика (в экономике)» / Н. И. Шанченко. – Ульяновск: УлГТУ, 2008. – 139 с.
7. Эконометрика: учеб. / под ред. И.И. Елисеевой. – М.: Проспект. 2009 – 288с.

Статистико-математические таблицы

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	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
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

k_2 \ k_1	1	2	3	4	5	6	8	12	24	∞
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

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

Число степеней свободы d.f.	**0,1**	**0,05**	**0,01**	**Число степеней свободы d.f.**	**0,1**	**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

Оглавление

Учебное издание

Трухина Людмила Ивановна

ПРАКТИКУМ ПО ЭКОНОМЕТРИКЕ:
ПАРНАЯ И МНОЖЕСТВЕННАЯ РЕГРЕССИЯ

Выпускающий редактор Е. И. Осянина
Подготовка оригинал-макета М. В. Голубцов

Подписано в печать 02.10.2023. Формат 60x84/16. Усл. печ. л. 5,28.
Тираж 100 экз. Заказ 1827.

Издательство «Бук». 420029, г. Казань, ул. Академика Кирпичникова, д. 25.
Отпечатано в издательстве «Бук»