Лекция 18

ҚАЗАҚСТАННЫҢ АШЫҚ УНИВЕРСИТЕТІ

ОСНОВЫ СТАТИСТИКИ

Регрессионный анализ. Парная регрессионная модель

В результате изучения материала вы освоите основы регрессионного анализа, научитесь рассчитывать параметры парной модели и давать оценку качества модели парной регрессии и корреляции.

Что же такое регрессионный анализ – это статистический метод исследования зависимости случайной величины от переменных. В статистическом моделировании регрессионный анализ представляет собой исследования, применяемые с целью оценки взаимосвязи между переменными. Этот математический метод включает в себя множество других методов для моделирования и анализа нескольких переменных, когда основное внимание уделяется взаимосвязи между зависимой переменной и одной или несколькими независимыми. Говоря более конкретно, регрессионный анализ помогает понять, как меняется типичное значение зависимой переменной, если одна из независимых переменных изменяется, в то время как другие независимые переменные остаются фиксированными. Во всех случаях целевая оценка является функцией независимых переменных и называется функцией регрессии. В регрессионном анализе также представляет интерес характеристика изменения зависимой переменной как функции регрессии, которая может быть описана с помощью распределения вероятностей.

Задачи регрессионного анализа.

Данный статистический метод исследования широко используется для прогнозирования, где его использование имеет существенное преимущество, но иногда это может приводить к иллюзии или ложным отношениям, поэтому рекомендуется аккуратно его использовать в указанном вопросе, поскольку, например, корреляция не означает причинно-следственной связи. Разработано большое число методов для проведения регрессионного анализа, такие как линейная и обычная регрессии по методу наименьших квадратов, которые являются параметрическими. Их суть в том, что функция регрессии определяется в терминах конечного числа неизвестных параметров, которые оцениваются из данных. Непараметрическая регрессия позволяет ее функции лежать в определенном наборе функций, которые могут быть бесконечномерными. Как статистический метод исследования, регрессионный анализ на практике зависит от формы процесса генерации данных и от того, как он относится к регрессионному подходу. Так как истинная форма процесса данных, генерирующих, как правило, неизвестное число, регрессионный анализ данных часто зависит в некоторой степени от предположений об этом процессе. Эти предположения иногда проверяемы, если имеется достаточное количество доступных данных. Регрессионные модели часто бывают, полезны даже тогда, когда предположения умеренно нарушены, хотя они не могут работать с максимальной эффективностью.

Применение регрессионного анализа.

Регрессионный анализ может использоваться в большом количестве приложений:

Моделирование числа поступивших в среднюю школу для лучшего понимания факторов, удерживающих детей в том же учебном заведении.

Моделирование дорожных аварий как функции скорости, дорожных условий, погоды и т.д., чтобы проинформировать полицию и снизить несчастные случаи.

Моделирование потерь от пожаров как функции от таких переменных как степень вовлеченности пожарных департаментов, время обработки вызова, или цена собственности. Если вы обнаружили, что время реагирования на вызов является ключевым фактором, возможно, существует необходимость создания новых пожарных станций. Если мы обнаружили, что вовлеченность – главный фактор, возможно, нам нужно увеличить оборудование и количество пожарных, отправляемых на пожар.

Термины регрессионного анализа.

Невозможно обсуждать регрессионный анализ без предварительного знакомства с основными терминами и концепциями, характерными для регрессионной статистики:

Уравнение регрессии. Это математическая формула, применяемая к независимым переменным, чтобы лучше спрогнозировать зависимую переменную, которую необходимо смоделировать. К сожалению, для тех ученых, кто думает, что x и y - это только координаты, независимая переменная в регрессионном анализе всегда обозначается как y, а зависимая – всегда X. Каждая независимая переменная связана с коэффициентами регрессии, описывающими силу и знак взаимосвязи между этими двумя переменными. Уравнение регрессии может выглядеть следующим образом (y – зависимая переменная, X – независимые

переменные, β – коэффициенты регрессии), ниже приводится описание каждого из этих компонентов уравнения регрессии):

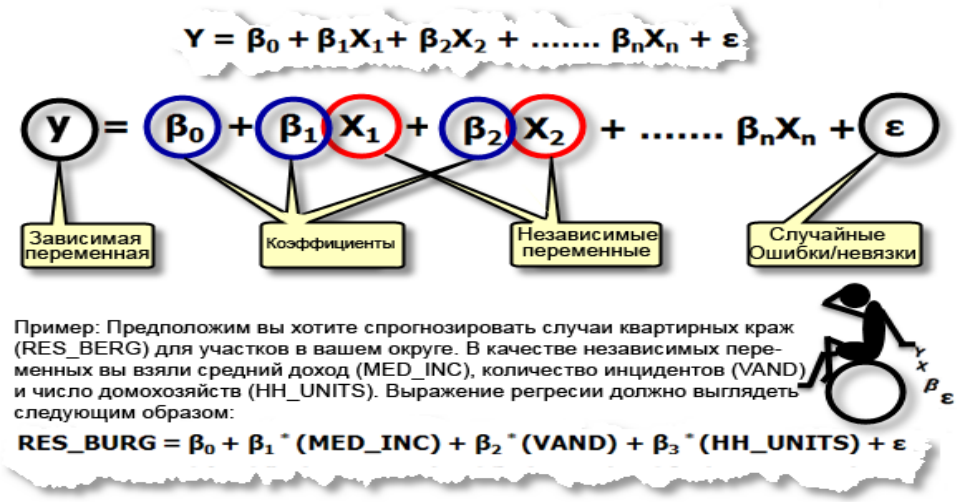

Рисунок 18.1 Элементы Уравнения регрессии по методу наименьших квадратов.

Зависимая переменная (y) – это переменная, описывающая процесс, который вы пытаетесь предсказать или понять (бытовые кражи, осадки). В уравнении регрессии эта переменная всегда находится слева от знака равенства. В то время, как можно использовать регрессию для предсказания зависимой величины, вы всегда начинаете с набора хорошо известных y-значений и используете их для калибровки регрессионной модели. Известные y-значения часто называют наблюдаемыми величинами.

Независимые переменные (X) – это переменные, используемые для моделирования или прогнозирования значений зависимых переменных. В уравнении регрессии они располагаются справа от знака равенства и часто называются независимыми переменными. Зависимая переменная – это функция независимых переменных. Если вас интересует прогнозирование годового оборота определенного магазина, можно включить в модель независимые переменные, отражающие, например, число потенциальных покупателей, расстояние до конкурирующих магазинов, заметность магазина и структуру спроса местных жителей.

Коэффициенты регрессии (β) – это коэффициенты, которые рассчитываются в результате выполнения регрессионного анализа. Вычисляются величины для каждой независимой переменной, которые представляют силу и тип взаимосвязи независимой переменной по отношению к зависимой. Предположим, что мы моделируем частоту пожаров как функцию от солнечной радиации, растительного покрова, осадков и экспозиции склона. Тогда можно ожидать положительную взаимосвязь между частотой пожаров и солнечной радиацией (другими словами, чем больше солнца, тем чаще встречаются пожары). Если отношение положительно, знак связанного коэффициента также положителен. Также можно ожидать негативную связь между частотой пожаров и осадками (другими словами, для мест с большим количеством осадков характерно меньше лесных пожаров). Коэффициенты отрицательных отношений имеют знак минуса. Когда взаимосвязь сильная, значения коэффициентов достаточно большие (относительно единиц независимой переменной, с которой они связаны). Слабая взаимосвязь описывается коэффициентами с величинами около 0; β0 – это пересечение линии регрессии. Он представляет ожидаемое значение зависимой величины, если все независимые переменные равны 0.

P-значения. Большинство регрессионных методов выполняют статистический тест для расчета вероятности, называемой p-значением, для коэффициентов, связанной с каждой независимой переменной. Нулевая гипотеза данного статистического теста предполагает, что коэффициент незначительно отличается от нуля (другими словами, для всех целей и задач, коэффициент равен нулю, и связанная независимая переменная не может объяснить нашу модель). Маленькие величины p-значений отражают маленькие вероятности и предполагают, что коэффициент действительно важен для нашей модели со значением, существенно отличающимся от 0 (другими словами, маленькие величины p-значений свидетельствуют о

том, что коэффициент не равен 0). Если коэффициент с p-значением, равным 0,01, например, статистически значимый для 99 % доверительного интервала, то связанные переменные являются эффективным предсказателем. Переменные с коэффициентами около 0 не помогают предсказать или смоделировать зависимые величины; они практически всегда удаляются из регрессионного уравнения, если только нет веских причин сохранить их.

R 2/R-квадрат: Статистические показатели составной R-квадрат и выровненный R-квадрат вычисляются из регрессионного уравнения, чтобы качественно оценить модель. Значение R-квадрат лежит в пределах от 0 до 100 процентов. Если наша модель описывает наблюдаемые зависимые переменные идеально, R-квадрат равен 1.0 (так как вы использовали модификацию величины y для предсказания y). Вероятнее всего, мы сможем наблюдать значения R-квадрат в районе 0,49, то можно, например, интерпретировать подобный результат как «Это модель объясняет 49 % вариации зависимой величины». Чтобы понять, как работает R-квадрат, постройте график, отражающий наблюдаемые и оцениваемые значения y, отсортированные по оцениваемым величинам. Обратим внимание на количество совпадений.

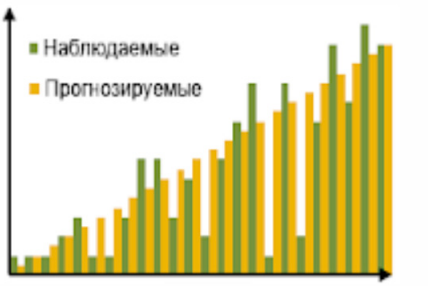

Рисунок 18.2 R-квадрат – это показатель моделирования, показывающий насколько хорошо оцененные y-значения совпадают с наблюдаемыми y-значениями.

Этот график визуально отображает, насколько хорошо вычисленные значения модели объясняют изменения наблюдаемых значений зависимых переменных. Выверенный R-квадрат всегда немного меньше, чем составной R-квадрат, т.к. он отражает всю сложность модели (количество переменных) и связан с набором исходных данных. Следовательно, выверенный R-квадрат является более точной мерой для оценки результатов работы модели.

Невязки. Существует необъяснимое количество зависимых величин, представленных в уравнении регрессии как случайные ошибки ε.

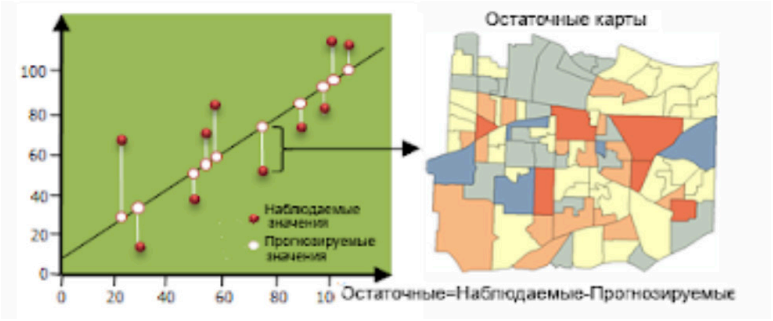

Рисунок 18.3 Невязки регрессионного анализа.

Красные области – местоположения, где реальные значения больше, нежели оцененные в модели. Синие области – местоположения, где реальные значения меньше, нежели оцененные моделью.

Известные значения зависимой переменной используются для построения и настройки модели регрессии. Используя известные величины зависимой переменной (Y) и известные значений для всех независимых переменных (Xs), регрессионный инструмент создаст уравнение, которое предскажет те известные y-значения как можно лучше. Однако предсказанные значения редко точно совпадают с наблюдаемыми величинами. Разница между наблюдаемыми и предсказываемыми значениями y называется

невязка или отклонение. Величина отклонений регрессионного уравнения – одно из измерений качества работы модели. Большие отклонения говорят о ненадлежащем качестве модели.

Рассмотрим простейшую модель парной регрессии – линейную регрессию. Линейная регрессия находит широкое применение в эконометрике ввиду четкой экономической интерпретации ее параметров.

Линейная регрессия сводится к нахождению уравнения вида

Формула 18.1 $y_x = a + b \cdot x$ или $y = a + b \cdot x + \varepsilon$.

Уравнение вида $y_x = a + b \cdot x$ позволяет по заданным значениям фактора x находить теоретические значения результативного признака, подставляя в него фактические значения фактора x.

Построение линейной регрессии сводится к оценке ее параметров – a и b. Классический подход к оцениванию параметров линейной регрессии основан на методе наименьших квадратов (МНК). МНК позволяет получить такие оценки параметров a и b, при которых сумма квадратов отклонений фактических значений результативного признака от теоретических минимальна:

Формула 18.2 $$\sum_{i=1}^{n}\left(y_i - \hat{y}_{x_i}\right)^2 = \sum_{i=1}^{n}\varepsilon_i^2 \rightarrow \min$$

Т.е. из всего множества линий линия регрессии на графике выбирается так, чтобы сумма квадратов расстояний по вертикали между точками и этой линией была бы минимальной (рис. 18.4):

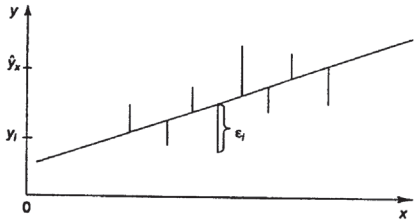

Рисунок 18.4 Линия регрессии с минимальной дисперсией остатков.

Найдем искомые оценки параметров a и b. Можно воспользоваться следующими готовыми формулами:

Формула 18.3 $a = \overline{y} - b \cdot \overline{x}$, $b = \dfrac{\operatorname{cov}(x, y)}{\sigma_x^2}$,

где $\operatorname{cov}(x,y) = \overline{y \cdot x} - \overline{y} \cdot \overline{x}$ – ковариация признаков x и y, $\sigma_x^2 = \overline{x^2} - \overline{x}^2$ – дисперсия признака x и

$\overline{x} = \dfrac{1}{n}\sum x$, $\overline{y} = \dfrac{1}{n}\sum y$, $\overline{y \cdot x} = \dfrac{1}{n}\sum y \cdot x$, $\overline{x^2} = \dfrac{1}{n}\sum x^2$.

Ковариация – числовая характеристика совместного распределения двух случайных величин, равная математическому ожиданию произведения отклонений этих случайных величин от их математических ожиданий. Дисперсия – характеристика случайной величины, определяемая как математическое ожидание квадрата отклонения случайной величины от ее математического ожидания. Математическое ожидание – сумма произведений значений случайной величины на соответствующие вероятности.

Параметр b называется коэффициентом регрессии. Его величина показывает среднее изменение результата с изменением фактора на одну единицу.

Возможность четкой экономической интерпретации коэффициента регрессии сделала линейное уравнение регрессии достаточно распространенным в эконометрических исследованиях.

Формально a – значение y при x=0. Если признак-фактор x не может иметь нулевого значения, то вышеуказанная трактовка свободного члена не имеет смысла, т.е. параметр a может не иметь экономического содержания.

Уравнение регрессии всегда дополняется показателем тесноты связи, о котором мы говорили ранее. Напомним, 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). Но следует иметь в виду, что близость абсолютной величины линейного коэффициента корреляции к нулю еще не означает отсутствия связи между признаками. При другой (нелинейной) спецификации модели связь между признаками может оказаться достаточно тесной.

Для оценки качества подбора линейной функции рассчитывается квадрат линейного коэффициента корреляции r^2_{xy}, называемый коэффициентом детерминации.

Формула 18.4

$$r^2_{xy} = 1 - \frac{\sigma^2_{ост}}{\sigma^2_y}$$

где $\sigma^2_{ост} = \frac{1}{n}\sum\left(y - \hat{y}_x\right)^2$, $\sigma^2_y = \frac{1}{n}\sum(y - \bar{y})^2 = \overline{y^2} - \bar{y}^2$

После того как найдено уравнение линейной регрессии, проводится оценка значимости как уравнения в целом, так и отдельных его параметров.

Проверить значимость уравнения регрессии – значит установить, соответствует ли математическая модель, выражающая зависимость между переменными, экспериментальным данным и достаточно ли включенных в уравнение объясняющих переменных (одной или нескольких) для описания зависимой переменной.

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

Формула 18.5

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

Средняя ошибка аппроксимации не должна превышать 8–10%.

Оценка значимости уравнения регрессии в целом производится на основе F-критерия Фишера, которому предшествует дисперсионный анализ. В математической статистике дисперсионный анализ рассматривается как самостоятельный инструмент статистического анализа. В эконометрике он применяется как вспомогательное средство для изучения качества регрессионной модели.

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

$$\sum(y - \bar{y})^2 = \sum\left(y_x - \bar{y}\right)^2 + \sum\left(y - \hat{y}_x\right)^2$$

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

$\sum\left(y_x - \bar{y}\right)^2$ – сумма квадратов отклонений, объясненная регрессией (или факторная сумма квадратов отклонений);

$\sum\left(y - \hat{y}_x\right)^2$ – остаточная сумма квадратов отклонений, характеризующая влияние неучтенных в модели факторов.

Схема дисперсионного анализа имеет вид, представленный в таблице 1.1 (n – число наблюдений, m – число параметров при переменной).

Таблица 18.1 Схема дисперсионного анализа

Компоненты дисперсии	Сумма квадратов	Число степеней свободы	Дисперсия на одну степень свободы
Общая	$\sum(y - \bar{y})^2$	n-1	$S^2_{общ} = \frac{\sum(y - \bar{y})^2}{n-1}$
Факторная	$\sum\left(y_x - \bar{y}\right)^2$	m	$S^2_{факт} = \frac{\sum\left(y_x - \bar{y}\right)^2}{m}$

Остаточная	$\sum\left(y-\hat{y}_x\right)^2$	n-m-1	$S^2_{ост}=\frac{\sum\left(y-\hat{y}_x\right)^2}{n-m-1}$

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:

Формула 18.6

$$F=\frac{S^2_{факт}}{S^2_{ост}}$$

Фактическое значение F-критерия Фишера (1.6) сравнивается с табличным значением $F_{табл}(\alpha; k_1; k_2)$ при уровне значимости α и степенях свободы k_1=m и k_2=n-m-1. При этом, если фактическое значение F-критерия больше табличного, то признается статистическая значимость уравнения в целом.

Для парной линейной регрессии m=1, поэтому

Формула 18.7

$$F=\frac{S^2_{факт}}{S^2_{ост}}=\frac{\sum\left(\hat{y}_x-\bar{y}\right)^2}{\sum\left(y-\hat{y}_x\right)^2}\cdot(n-2)$$

Формула 18.7

Величина F-критерия связана с коэффициентом детерминации r^2_{xy}, и ее можно рассчитать по следующей формуле:

Формула 18.8

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

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

Стандартная ошибка коэффициента регрессии определяется по формуле:

Формула 18.

$$m_b=\sqrt{\frac{S^2_{ост}}{\sum\left(x-\bar{x}\right)^2}}=\frac{S_{ост}}{\sigma_x\cdot\sqrt{n}}$$

где $S^2_{ост}=\frac{\sum\left(y-\hat{y}_x\right)^2}{n-2}$ – остаточная дисперсия на одну степень свободы.

Величина стандартной ошибки совместно с t-распределением Стьюдента при n-2 степенях свободы применяется для проверки существенности коэффициента регрессии и для расчета его доверительного интервала.

Для оценки существенности коэффициента регрессии его величина сравнивается с его стандартной ошибкой, т.е. определяется фактическое значение t-критерия Стьюдента: $t_b=\frac{b}{m_b}$ которое затем сравнивается с табличным значением при определенном уровне значимости и числе степеней свободы (n-2). Доверительный интервал для коэффициента регрессии определяется как $b\pm t_{табл}\cdot m_b$. Поскольку знак коэффициента регрессии указывает на рост результативного признака у при увеличении признака-фактора x(b>0), уменьшение результативного признака при увеличении признака-фактора () или его независимость от независимой переменной (b<0) (см. рис. 18.5), то границы доверительного интервала для коэффициента регрессии не должны содержать противоречивых результатов, например, $-1{,}5\le b\le 0{,}8$. Такого рода запись указывает, что истинное значение коэффициента регрессии одновременно содержит положительные и

отрицательные величины и даже ноль, чего не может быть.

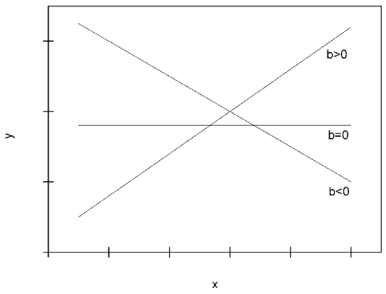

Рисунок 18.5. Наклон линии регрессии в зависимости от значения параметра b.

Стандартная ошибка параметра определяется по формуле:

Формула 18.10

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

Процедура оценивания существенности данного параметра не отличается от рассмотренной выше для коэффициента регрессии. Вычисляется t-критерий: $t_a = \frac{a}{m_a}$, его величина сравнивается с табличным значением при n-2 степенях свободы.

Значимость линейного коэффициента корреляции проверяется на основе величины ошибки коэффициента корреляции m_2:

Формула 18.11

$$m_r = \sqrt{\frac{1 - r^2}{n - 2}}$$

Фактическое значение -критерия Стьюдента определяется как $t_r = \frac{r}{m_r}$.

Существует связь между -критерием Стьюдента и -критерием Фишера:

Формула 18.12

$$t_b = t_r = \sqrt{F}$$

В прогнозных расчетах по уравнению регрессии определяется предсказываемое y_p значение как точечный прогноз y_x при $x_p = x_k$, т.е. путем подстановки в уравнение регрессии $y_x = a + b \cdot x$ соответствующего значения x. Однако точечный прогноз явно нереален. Поэтому он дополняется расчетом стандартной ошибки y_p, т.е. m_{y_p} , и соответственно интервальной оценкой прогнозного значения y_p:

$$y_p - \Delta_{y_p} \leq \hat{y}_p \leq \hat{y}_p + \Delta_{y_p}$$

где $\Delta_{y_p} = m_{y_p} \cdot t_{\text{табл}}$, а m_{y_p} – средняя ошибка прогнозируемого индивидуального значения:

Формула 18.13

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

Рассмотрим пример. По данным проведенного опроса восьми групп семей известны данные связи расходов населения на продукты питания с уровнем доходов семьи.

Таблица 18.2 Исходные данные

Расходы на продукты питания, у млн. тенге	0,9	1,2	1,8	2,2	2,6	2,9	3,3	3,8

Доходы семьи, x, млн. тенге	1,2	3,1	5,3	7,4	9,6	11,8	14,5	18,7

Предположим, что связь между доходами семьи и расходами на продукты питания линейная. Для подтверждения нашего предположения построим поле корреляции.

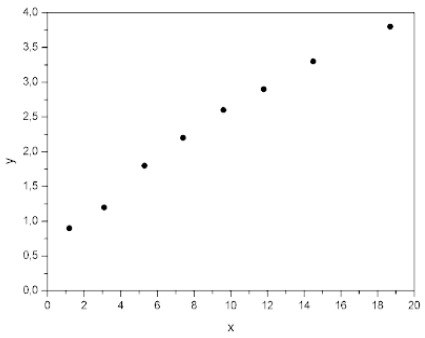

Рисунок 18.6 Поле корреляции

По графику видно, что точки выстраиваются в некоторую прямую линию.

Для удобства дальнейших вычислений составим таблицу 18.1.

Таблица 18.3 Таблица вычислений

	x	y	x*y	x^2	y^2	y_x	$y-\hat{y}_x$	$(y-\hat{y}_x)^2$	A_i, %
1	2	3	4	5	6	7	8	9	10
1	1,2	0,9	1,08	1,44	0,81	1,038	– 0,138	0,0190	15,33
2	3,1	1,2	3,72	9,61	1,44	1,357	– 0,157	0,0246	13,08
3	5,3	1,8	9,54	28,09	3,24	1,726	0,074	0,0055	4,11
4	7,4	2,2	16,28	54,76	4,84	2,079	0,121	0,0146	5,50
5	9,6	2,6	24,96	92,16	6,76	2,449	0,151	0,0228	5,81
6	11,8	2,9	34,22	139,24	8,41	2,818	0,082	0,0067	2,83
7	14,5	3,3	47,85	210,25	10,89	3,272	0,028	0,0008	0,85
8	18,7	3,8	71,06	349,69	14,44	3,978	– 0,178	0,0317	4,68
Итого	71,6	18,7	208,71	885,24	50,83	18,717	– 0,017	0,1257	52,19
Среднее значение	8,95	2,34	26,09	110,66	6,35	2,34	–	0,0157	6,52
σ	5,53	0,935	–	–	–	–	–	–	–
σ^2	30,56	0,874	–	–	–	–	–	–	–

Рассчитаем параметры линейного уравнения парной регрессии $y_x = a + b \cdot x$. Для этого воспользуемся формулами (18.5):

$$b = \frac{\mathrm{cov}(x, y)}{\sigma_x^2} = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{26,09 - 8,95 \cdot 2,34}{30,56} = 0,168$$

$$a = \overline{y} - b \cdot \overline{x} = 2,34 - 0,168 \cdot 8,95 = 0,836$$

Получили уравнение: y_x=0,836+0,168*x. Т.е. с увеличением дохода семьи на 1000 тенге расходы на питание увеличиваются на 168 тенге

Как было указано выше, уравнение линейной регрессии всегда дополняется показателем тесноты связи – линейным коэффициентом корреляции r_{xy}:

$$r_{xy} = b \cdot \frac{\sigma_x}{\sigma_y} = 0,168 \cdot \frac{5,53}{0,935} = 0,994$$

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

Коэффициент детерминации r^2_{xy}=0,987 показывает, что уравнением регрессии объясняется 98,7% дисперсии результативного признака, а на долю прочих факторов приходится лишь 1,3%.

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

$$F = \frac{r_{xy}^2}{1-r_{xy}^2}\cdot(n-2) = \frac{0,987}{1-0,987}\cdot 6 = 455,54$$

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

Для оценки статистической значимости коэффициентов регрессии и корреляции рассчитаем t-критерий Стьюдента и доверительные интервалы каждого из показателей. Рассчитаем случайные ошибки параметров линейной регрессии и коэффициента корреляции

$$\left(S^2_{ост} = \frac{\sum\left(y-\hat{y}_x\right)^2}{n-2} = \frac{0,1257}{8-2} = 0,021\right)$$

$$m_b = \frac{S_{ост}}{\sigma_x\cdot\sqrt{n}} = \frac{\sqrt{0,021}}{5,53\cdot\sqrt{8}} = 0,0093,$$

$$m_a = S_{ост}\cdot\frac{\sqrt{\sum x^2}}{\sigma_x\cdot n} = \frac{\sqrt{0,021\cdot 885,24}}{5,53\cdot 8} = 0,0975,$$

$$m_r = \sqrt{\frac{1-r^2}{n-2}} = \sqrt{\frac{1-0,987}{6}} = 0,0465.$$

Фактические значения t-статистик: $t_b = \frac{0,168}{0,0093} = 18,065$ $t_a = \frac{0,836}{0,0975} = 8,574$ $t_r = \frac{0,994}{0,0465} = 21,376$ Табличное значение t-критерия Стьюдента при α=0,05 и числе степеней свободы v=n-2=6 есть $t_{табл}$=2,447. Так как t_b>$t_{табл}$, t_a>$t_{табл}$ и t_r> $t_{табл}$, то признаем статистическую значимость параметров регрессии и показателя тесноты связи. Рассчитаем доверительные интервалы для параметров регрессии a и b: $a \pm t\cdot m_a$ и $b \pm t\cdot m_b$. Получим, что $a \in [0,597;\ 1,075]$ и $b \in [0,145;\ 0,191]$.

Средняя ошибка аппроксимации (находим с помощью столбца 10 таблицы 18.3; $A_i = \left|\frac{y_i-\hat{y}_{x_i}}{y_i}\right|\cdot 100\%$) $\bar{A}$=6,52% говорит о хорошем качестве уравнения регрессии, т.е. свидетельствует о хорошем подборе модели к исходным данным.

И, наконец, найдем прогнозное значение результативного фактора y_p при значении признака-фактора, составляющем 110% от среднего уровня $x_p = 1,1\cdot\bar{x} = 1,1\cdot 8,95 = 9,845$, т.е. найдем расходы на питание, если доходы семьи составят 9,85 млн. тенге

y_p=0,836+0,168*9,845=2,490 (млн. тенге)

Значит, если доходы семьи составят 9,845 м. млн. тенге, то расходы на питание будут 2,490 млн. тенге

Найдем доверительный интервал прогноза. Ошибка прогноза

$$m_{y_p} = S_{ост}\cdot\sqrt{1+\frac{1}{n}+\frac{\left(x_p-\bar{x}\right)^2}{n\cdot\sigma_x^2}} = \sqrt{0,021\cdot\left(1+\frac{1}{8}+\frac{(9,845-8,95)^2}{8\cdot 30,56}\right)} = 0,154$$

а доверительный интервал ($\hat{y}_p - \Delta_{y_n} \le \hat{y}_p \le \hat{y}_p + \Delta_{y_n}$):

$2,113<y_p<2,867$.

т.е. прогноз является статистически надежным.

Теперь на одном графике изобразим исходные данные и линию регрессии:

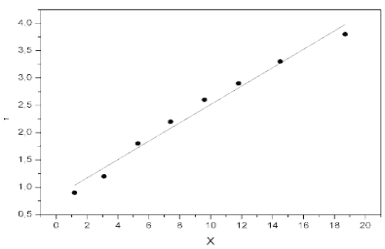

Рисунок 18.7 Исходные данные и линия регрессии

Выводы:

Парный коэффициент корреляции показывает тесноту связи между показателями, существует прямая и обратная связь.

Коэффициент регрессии показывает насколько изменится результат при изменении независимого фактора на одну единицу.

Допустимая ошибка аппроксимации составляет 8%-10 % и показывает качество подмора модели по отношению к исходным данным.

Регрессионная модель позволяет троить прогнозные значения.