ПАРНАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

Парная линейная регрессия – регрессионная зависимость между двумя переменными y и x, т. е. модель вида $y = a + bx + e$, где y – отклик, x – фактор, e - случайная «остаточная» компонента.

Далее рассмотрим алгоритм вычисления для случая парной линейной регрессии.

1. **Выбор формы модели**

   На этом этапе по исходным статистическим данным выбирается наиболее подходящая модель, т. е. уравнение регрессии. Этот выбор может быть осуществлен тремя способами:

   - Графическим
     На координатной плоскости строят точки с координатами (x, y), по расположению которых предполагают наличие зависимости определенного вида. Такое изображение статистической зависимости называется *диаграммой рассеяния*.
   - Аналитическим, т. е. исходя из теории изучаемой взаимосвязи
   - Экспериментальным

2. **Вычисление коэффициентов (параметров) регрессионной модели**

   Для оценки параметров используется *метод наименьших квадратов* (МНК), согласно которому неизвестные параметры a и b выбираются таким образом, чтобы сумма квадратов отклонений фактических значений отклика y от прогнозных (полученных по уравнению регрессии) y была минимальна, т. е.

$$\sum \left(y - y \right)^2 \to \min .$$

   Чтобы найти минимум функции, надо вычислить частные производные по каждому из параметров a и b и приравнять их к нулю. Обозначим: $\sum \left(y - y \right)^2 = \sum \left(y - a - bx \right)^2 = S(a,b)$. Тогда,

$$\begin{cases} \dfrac{\partial S}{\partial a} = \sum (2(y - a - bx)(-1)) = 0 \\ \dfrac{\partial S}{\partial b} = \sum (2(y - a - bx)(-x)) = 0 \end{cases} \Rightarrow \begin{cases} \sum (-2y + 2a + 2bx) = 0 \\ \sum (-2yx + 2ax + 2bx^2) = 0 \end{cases} \Rightarrow$$

$$\Rightarrow \begin{cases} -2\sum y + 2na + 2b\sum x = 0 \\ -2\sum yx + 2a\sum x + 2b\sum x^2 = 0 \end{cases} \Rightarrow \begin{cases} Na + b\sum x = \sum y \\ a\sum x + b\sum x^2 = \sum yx \end{cases}$$

   Делим обе части уравнений системы на N (объем выборки). Получим:

$$\begin{cases} a + b\overline{x} = \overline{y} \\ a\overline{x} + b\overline{x^2} = \overline{xy} \end{cases} \Rightarrow \boxed{\begin{cases} a = \overline{y} - b\overline{x} \\ b = \dfrac{\overline{yx} - \overline{y} \cdot \overline{x}}{\overline{x^2} - \left(\overline{x}\right)^2} \end{cases}}.$$

   Формально a – значение y при $x = 0$. Если фактор x не имеет и не может иметь нулевого значения, то такая трактовка параметра a не имеет смысла. Этот параметр может и не иметь экономического содержания. Попытки экономически интерпретировать параметр a могут привести к абсурду, особенно при $a < 0$.

   Интерпретировать можно лишь знак при параметре a. *Если* $a > 0$*, то относительное* *изменение отклика происходит медленнее, чем изменение фактора.*

Коэффициент b называется ***коэффициентом регрессии*** и *показывает среднее изменение отклика при изменении фактора на одну единицу.*

3. **Оценка значимости коэффициента регрессии**

Используется *t-критерий Стьюдента*. Выдвигается гипотеза $H_0 : b = 0$ об отсутствии влияния фактора на отклик. Вычисляется фактическое значения *t*-критерия:

$$t_b = \frac{|b|}{SE_b},$$

где SE_b - *стандартная ошибка*, вычисляемая по формуле:

$$SE_b = \sqrt{\frac{\sum\left(y - y\right)^2}{(N-1)\cdot(N-2)\cdot\sigma_x^2}} = \left|\text{остатки } e_i = y_i - y_i\right| = \sqrt{\frac{\sum e_i^2}{(N-1)\cdot(N-2)\cdot\sigma_x^2}}.$$

Стандартные отклонения для фактора и отклика вычисляются по формулам:

$$\sigma_x = \sqrt{\frac{\sum\left(x_i - \overline{x}\right)^2}{N-1}} = \sqrt{\frac{N}{N-1}\cdot\left(\overline{x^2} - \left(\overline{x}\right)^2\right)} \text{ и } \sigma_y = \sqrt{\frac{\sum\left(y_i - \overline{y}\right)^2}{N-1}} = \sqrt{\frac{N}{N-1}\cdot\left(\overline{y^2} - \left(\overline{y}\right)^2\right)}.$$

Сравнивая фактическое и табличное $t_{табл}$ значения на уровне значимости $\alpha = 0,05$ и числе степеней свободы $N-2$ принимается решение:

- Если $t_{табл} < t_b$, то H_0 отклоняется, влияние фактора на отклик обнаружено.
- Если $t_{табл} > t_b$, то гипотеза H_0 принимается, фактор на отклик не оказывает существенного влияния.

Далее необходимо рассчитать ***границы доверительного интервала для коэффициента регрессии***. Вычисленное выше значение коэффициента регрессии – это среднее значение. Доверительный интервал показывает, каким может быть коэффициент регрессии в 95% случаев. Формулы для расчета имеют следующий вид:

$$н.гр. = b - t_{табл}\cdot SE_b,\ в.гр. = b + t_{табл}\cdot SE_b.$$

Рассчитывать границы доверительного интервала следует лишь тогда, когда влияние фактора на отклик обнаружено.

4. **Оценка качества всей модели**

Тесноту связи отклика и фактора оценивает ***линейный коэффициент корреляции Пирсона*** r, который можно вычислить, например, по следующей формуле: $r = b\frac{\sigma_x}{\sigma_y}$.

Важным свойством коэффициента корреляции является следующее: $-1 \le r \le 1$. Если $r > 0$, то корреляционная связь между переменными называется *прямой*, если $r < 0$ – *обратной*. Качественная оценка тесноты связи может быть выявлена на основе *шкалы Чеддока*.

Теснота связи	Значение $\lvert r\rvert$
слабая	0,1 – 0,3
умеренная	0,3 – 0,5
заметная	0,5 – 0,7
высокая	0,7 – 0,9
весьма высокая	0,9 – 0,99

Коэффициент детерминации – для парной линейной регрессии – это квадрат линейного коэффициента корреляции. Величина R^2 показывает, сколько процентов отклика объясняется с помощью включенного в модель фактора. Чем больше R^2, тем лучше построенная модель. Если $R^2 < 30\%$, то прогнозировать по такой модели нецелесообразно.

Для оценивания качества уравнения регрессии используется *F-критерий Фишера*, который состоит в проверке гипотезы H_0 о статистической незначимости уравнения регрессии и коэффициента детерминации $\left(R^2 = 0\right)$. Для этого необходимо вычислить фактическое значение F-критерия:

$$F_{факт} = \frac{R^2}{1-R^2}(N-2).$$

Сравнивая фактическое значение $F_{факт}$ и табличное $F_{табл}$ на уровне значимости $\alpha = 0,05$ и числе степеней свободы 1 и $N-2$ принимается решение:

- Если $F_{табл} < F_{факт}$, то гипотеза H_0 отклоняется, полученное уравнение значимо, т.е. построенная модель «лучше» прогноза по среднему.
- Если $F_{табл} > F_{факт}$, то гипотеза H_0 принимается, уравнение незначимо, качество построенной модели сравнимо с точностью прогноза по среднему.

5. Анализ остатков

Для проверки целесообразности использования линейной регрессионной модели используется процедура графического анализа остатков. ***Остатки должны быть нормально распределены и не зависеть от предсказанных по уравнению регрессии значений отклика***. Для такой проверки по столбцу остатков строится гистограмма и ящик с усами – для проверки на нормальность и диаграмма рассеяния для проверки независимости остатков от прогноза.

Алгоритм оценки остатков:

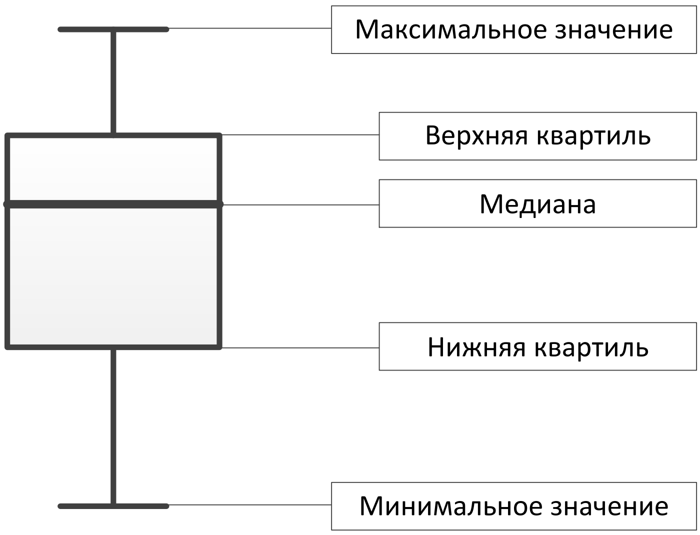
1) по формуле Стерджесса определить количество интервалов, на которые следует разбить остатки: $n = 1 + 3,322 \cdot \lg N$;
2) определить длину интервала: $h = \dfrac{e_{\max} - e_{\min}}{n-1}$;
3) найти границы первого интервала: $\left[e_{\min} - \dfrac{h}{2}; e_{\min} + \dfrac{h}{2}\right]$;
4) рассчитать границы остальных интервалов и частоты (количество остатков, попавших в тот или иной интервал). Получим интервальный вариационный ряд;
5) построить гистограмму. В случае нормального распределения остатков гистограмма будет иметь (почти) симметричный вид;
6) упорядочить столбик остатков по возрастанию;
7) по полученным данным найти медиану, верхнюю и нижнюю квартили, минимальное и максимальное знаечния:

   Медиана – значение в упорядоченной выборке, которое делит ее на две равные части.

   Нижняя квартиль – значение в упорядоченной выборке, которое отделяет 25%, начиная с минимального.

   Верхняя квартиль – значение в упорядоченной выборке, которое отделяет 25%, начиная с максимального.

8) построить ящик с усами по следующей схеме:



В случае нормального распределения ящик будет (почти) симметричным.

9) построить диаграмму рассеяния, на которой изобразить точки с координатами $\left(y_i, e_i\right)$. В случае «хорошей» модели никакой закономерности в расположении точек не должно прослеживаться.

6. Оценка точности прогнозов и прогнозирование

Средняя абсолютная ошибка модели (Mean Absolute Percent Error – MAPE) показывает, на сколько процентов в среднем прогноз отклоняется от факта. Вычисляется по формуле:

$$MAPE = \frac{\sum \left| \frac{e_i}{y_i} \right|}{N} \cdot 100\% .$$

Показатель MAPE полезен при использовании разных методов построения модели на основе одних и тех же данных. Чем меньше MAPE, тем лучше модель.

Прогнозное значение y_f определяется путем подстановки в уравнение линейной регрессии соответствующего прогнозного значения x_f. Далее вычисляется стандартная ошибка прогноза:

$$SE_f = \sqrt{\frac{\sum e^2}{N-2} \cdot \left(1 + \frac{1}{N} + \frac{(x_f - \bar{x})^2}{\sigma_x^2 \cdot \left(N-1\right)}\right)} .$$

95-процентный доверительный интервал прогноза строится по формуле:

$$н.гр. = y_f - t_{табл} \cdot SE_f ,\ в.гр. = y_f + t_{табл} \cdot SE_f .$$

Пример 1

Постановка задачи

По территориям региона за 199Хг. имеются следующие данные:

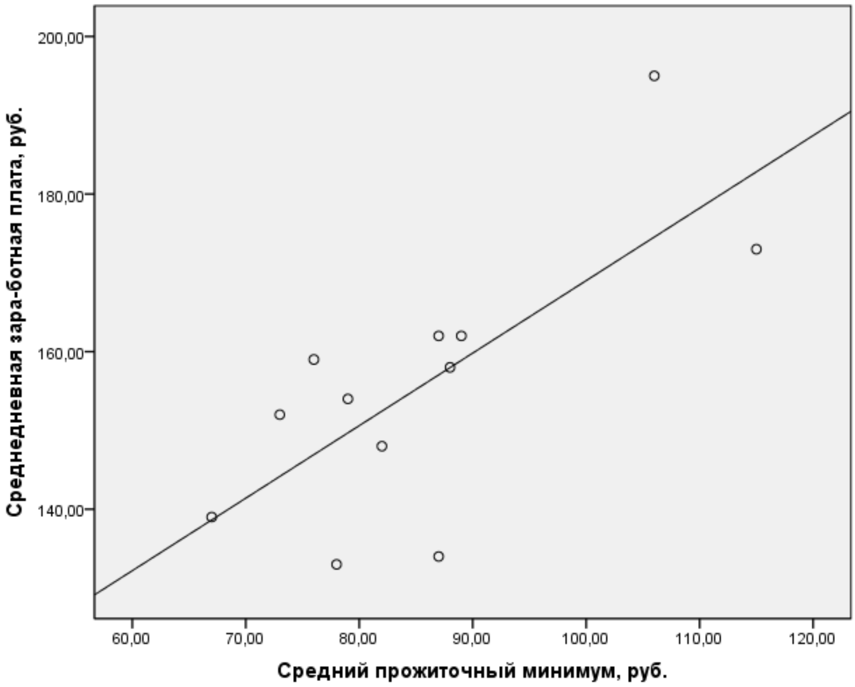
Номер региона	Средний прожиточный минимум, *x*, руб.	Среднедневная заработная плата, *y*, руб.
1	78	133
2	82	148
3	87	134
4	79	154
5	89	162
6	106	195
7	67	139
8	88	158
9	73	152
10	87	162
11	76	159
12	115	173

Задание

1. Изобразить диаграмму рассеяния и сформулировать гипотезу о форме связи.
2. Найти параметры a и b уравнения парной линейной регрессии $y = a + bx$. Пояснить эконометрический смысл параметра b.
3. Оценить статистическую значимость коэффициента регрессии (b) используя t-критерий Стьюдента на уровне значимости $\alpha = 0,05$.
4. Рассчитать границы доверительного интервала для параметра b.
5. Вычислить коэффициент корреляции r и оценить тесноту связи между фактором и откликом.
6. Вычислить коэффициент детерминации R^2, пояснить его эконометрический смысл и проверить его значимость с использованием F-критерия Фишера при $\alpha = 0,05$.
7. Проанализировать остатки.
8. Вычислить MAPE.
9. Выполнить прогноз среднедневной заработной платы при прогнозном значении прожиточного минимума 83 руб. Оценить доверительный интервал прогноза.

<u>Решение</u>

1. **Графический анализ** – построение диаграммы рассеяния, по которой определяется форма регрессионной модели.



По расположению точек предположим наличие линейной зависимости $y = a + bx$.

2. **Вычислим параметры** *a* и *b* уравнения парной линейной регрессии $y = a + bx$.

Для расчета параметров уравнения линейной регрессии составляем расчетную таблицу. Сначала заполняем столбцы с (1) по (5).

№	x	y	xy	x^2	y^2	y	$e = y - y$	e^2	$\left\|\frac{e}{y}\right\| \cdot 100\%$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	78	133	10374	6084	17689	149	-16	256	12,03
2	82	148	12136	6724	21904	152	-4	16	2,70
3	87	134	11658	7569	17956	157	-23	529	17,16
4	79	154	12166	6241	23716	150	4	16	2,60
5	89	162	14418	7921	26244	159	3	9	1,85
6	106	195	20670	11236	38025	175	20	400	10,26
7	67	139	9313	4489	19321	139	0	0	0,00
8	88	158	13904	7744	24964	158	0	0	0,00
9	73	152	11096	5329	23104	144	8	64	5,26
10	87	162	14094	7569	26244	157	5	25	3,09
11	76	159	12084	5776	25281	147	12	144	7,55
12	115	173	19895	13225	29929	183	-10	100	5,78
Сумма	1027	1869	161808	89907	294377			1559	68,28
Среднее	85,58	155,75	13484	7492,25	24531,42				5,69

$$b = \frac{\overline{yx} - \overline{y} \cdot \overline{x}}{\overline{x^2} - \left(\overline{x}\right)^2} = \frac{13484 - 85,58 \cdot 155,75}{7492,25 - 85,58^2} = 0,92;$$

$$a = \overline{y} - b \cdot \overline{x} = 155,75 - 0,92 \cdot 85,58 = 77,02.$$

Получено уравнение регрессии: $y = 77,02 + 0,92x$.

Эконометрический смысл коэффициента регрессии: с увеличением среднего прожиточного минимума на 1 руб. среднедневная заработная плата возрастает в среднем на 0,92 руб.

3. **Проверим статистическую значимость** коэффициента регрессии.

   Используем *t*-критерий Стьюдента. Выдвигаем гипотезу $H_0 : b = 0$ об отсутствии влияния фактора на отклик. Далее, необходимо сначала заполнить столбцы с (6) по (8), затем вычислить стандартную ошибку:

$$SE_b = \sqrt{\frac{\sum e_i^2}{(N-1)\cdot(N-2)\cdot\sigma_x^2}} = \left|\sigma_x = \sqrt{\frac{N}{N-1}\cdot\left(\overline{x^2} - \left(\overline{x}\right)^2\right)} = \sqrt{\frac{12}{12-1}\cdot\left(7492,25 - 85,58^2\right)} = 13,55\right| =$$

$$= \sqrt{\frac{1559}{(12-1)\cdot(12-2)\cdot 13,55^2}} = 0,278.$$

   Фактическое значение *t*-критерия Стьюдента: $t_b = \dfrac{|b|}{SE_b} = \dfrac{0,92}{0,278} = 3,309$.

   $t_{табл}$ на уровне значимости $\alpha = 0,05$ и числа степеней свободы $N - 2 = 12 - 2 = 10$ равно 2,228.

   $t_b = 3,309 > t_{табл} = 2,228$, гипотеза H_0 отклоняется, т. е. влияние фактора на отклик обнаружено.

4. **Границы 95-процентного доверительного интервала** для коэффициента регрессии:

$$н.гр. = b - t_{табл}\cdot SE_b = 0,92 - 2,228\cdot 0,278 = 0,301,$$
$$в.гр. = b + t_{табл}\cdot SE_b = 0,92 + 2,228\cdot 0,278 = 1,539.$$

   При увеличении среднего прожиточного минимума на 1 руб. среднедневная заработная плата вырастет в среднем на 0,92 руб., в 95% случаев рост может составлять от 0,3 руб. до 1,5 руб.

5. **Вычислим коэффициент корреляции**:

$$r = b\frac{\sigma_x}{\sigma_y} = \left|\sigma_y = \sqrt{\frac{N}{N-1}\cdot\left(\overline{y^2} - \left(\overline{y}\right)^2\right)} = \sqrt{\frac{12}{12-1}\cdot\left(24531,42 - 155,75^2\right)} = 17,27\right| =$$

$$= 0,92\cdot\frac{13,55}{17,27} = 0,721.$$

   Корреляция больше нуля, значит связь прямая, по шкале Чеддока – заметная.

6. **Вычислим коэффициент детерминации**: $R^2 = 0,721^2 = 0,520$. Это означает, что 52% заработной платы объясняется с помощью фактора «средний прожиточный минимум». $R^2 = 52\% > 30\%$, значит прогнозировать по данной модели целесообразно.

   Проверим статистическую значимость уравнения регрессии с помощью *F*-критерия Фишера. Выдвигаем гипотезу H_0 о статистической незначимости уравнения регрессии и коэффициента детерминации. Фактическое значение *F*-критерия равно:

$$F_{факт} = \frac{R^2}{1-R^2}(N-2) = \frac{0,520}{1-0,520}(12-2) = 10,83.$$

   $F_{табл} = 4,694$ на уровне значимости $\alpha = 0,05$ и числе степеней свободы 1 и

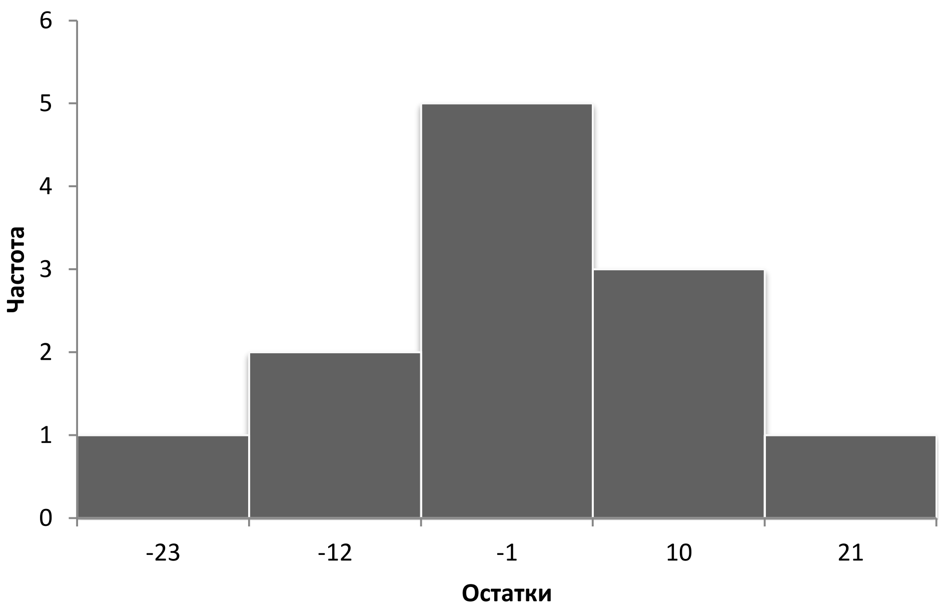
$N-2=12-2=10$.

$F_{факт}=10,83>F_{табл}=4,964$, гипотеза H_0 отклоняется и признается статистическая значимость уравнения регрессии. Построенная модель «лучше» прогноза по среднему.

7. **Проанализируем остатки**.
Проверим остатки на нормальность графическим способом:
1) Изобразим гистограмму. Для этого построим интервальный вариационный ряд. Число интервалов, на которые разобьем найденные остатки, определим по формуле Стерджесса
$n=1+3,322\cdot\lg N=1+3,322\cdot\lg 12\approx 5$.
Длина интервала: $h=\dfrac{R}{n-1}=\dfrac{e_{\max}-e_{\min}}{n-1}=\dfrac{20-(-23)}{5-1}=11$.
Границы первого интервала определим следующим образом:
$\left[e_{\min}-\dfrac{h}{2};e_{\min}+\dfrac{h}{2}\right]=\left[-23-\dfrac{11}{2};-23+\dfrac{11}{2}\right]=[-28,5;-17,5]$.

№	Нижняя граница	Верхняя граница	Частота
1	-28,5	-17,5	1
2	-17,5	-6,5	2
3	-6,5	4,5	5
4	4,5	15,5	3
5	15,5	26,5	1
Сумма			12

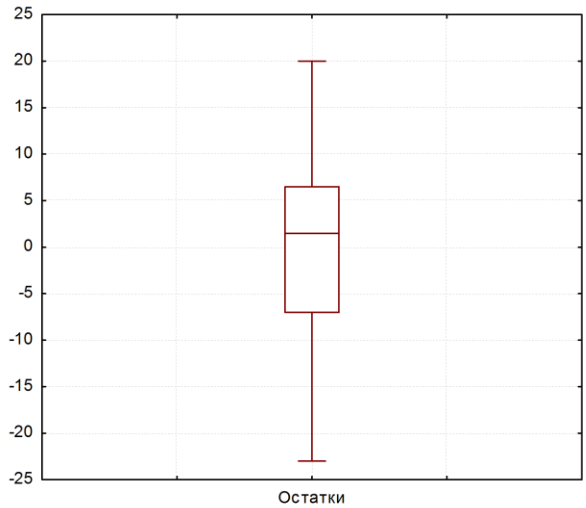
Гистограмма будет иметь вид:



2) Изобразим ящик с усами. Для этого упорядочим остатки по возрастанию и найдем медиану и квартили.

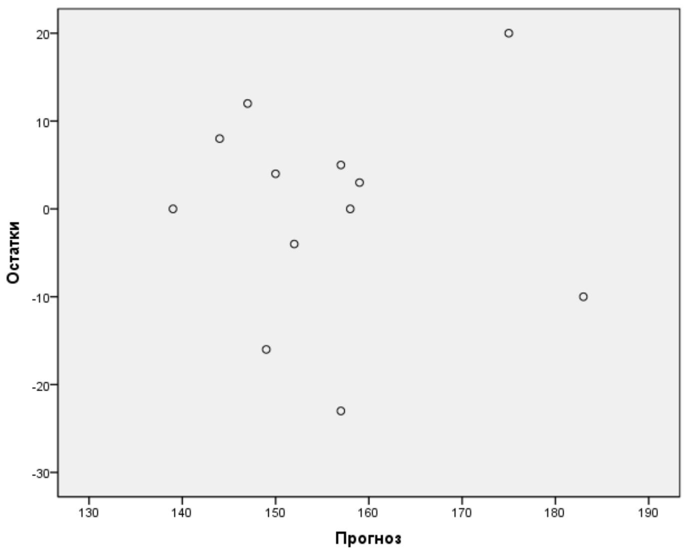
Остатки	Упорядоченные остатки	
-16	-23	минимум
-4	-16	
-23	-10	$Q_н = \frac{-10-4}{2} = -7$
4	-4	
3	0	
20	0	$Me = \frac{0+3}{2} = 1,5$
0	3	
0	4	
8	5	$Q_в = \frac{5+8}{2} = 6,5$
5	8	
12	12	
-10	20	максимум

Ящик с усами будет иметь вид:



И гистограмма, и ящик с усами являются симметричными, что свидетельствует о нормальном распределении остатков.

С помощью диаграммы рассеяния проверим, чтобы остатки не зависели от предсказанных по уравнению регрессии значений.



Остатки не зависят от предсказанных по уравнению регрессии значений.

Все это говорит о приемлемости линейной модели для описания скрытой в исходных данных зависимости.

8. **Оценим точность прогнозов**, вычислим среднюю абсолютную ошибку прогноза в процентах (MAPE). Заполнив столбец (9) в расчетной таблице, получим $MAPE = 5,69\%$. Это означает, что при прогнозе по построенной модели ошибка в среднем будет составлять 5,69%. Такой результат будем считать приемлемым.
9. **Вычислим прогнозное значение отклика** y_f, если прогнозное значение фактора $x_f = 83$ руб.

   $y_f = 77,02 + 0,92 \cdot 83 \cong 153.$ руб.

   Стандартная ошибка прогноза:

$$SE_f = \sqrt{\frac{\sum e^2}{N-2} \cdot \left(1 + \frac{1}{N} + \frac{(x_f - \overline{x})^2}{\sigma_x^2 \cdot (N-1)}\right)} = \sqrt{\frac{1559}{12-2} \cdot \left(1 + \frac{1}{12} + \frac{(83-85,58)^2}{13,55^2 \cdot (12-1)}\right)} = 13,016 .$$

   95-процентный доверительный интервал прогноза строится по формуле:

$$н.гр. = y_f - t_{табл} \cdot SE_f = 153 - 2,228 \cdot 13,016 = 124,$$

$$в.гр. = y_f + t_{табл} \cdot SE_f = 153 + 2,228 \cdot 13,016 = 182.$$

   Для прожиточного минимума, равного 83 руб. значение заработной платы с вероятностью 0,95 попадет в интервал от 124 руб. до 182 руб.

Пример 2

Постановка задачи

Имеются следующие данные о покупателях продуктовых магазинов некоторой сети:

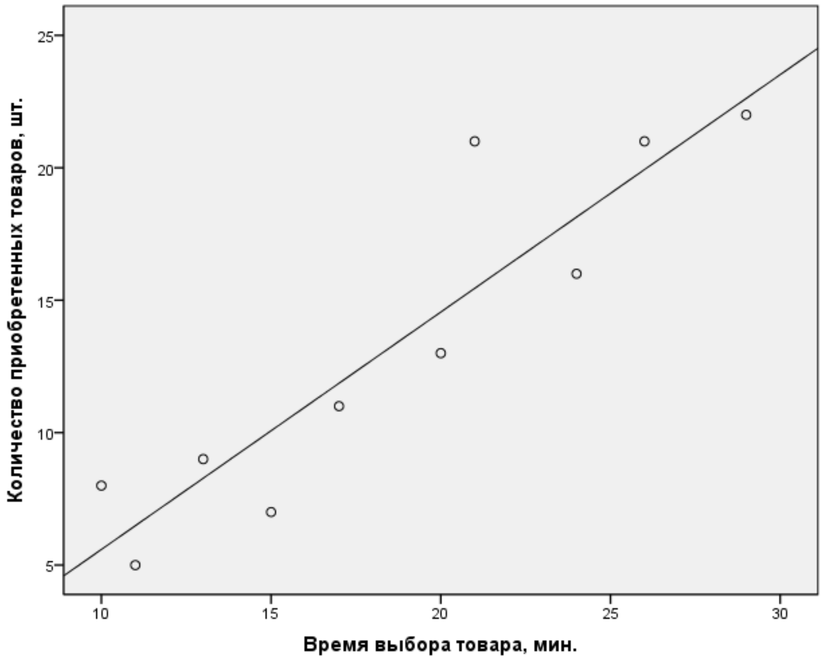
№ покупателя	Количество приобретенных товаров, шт., y	Время выбора товара, мин., x
1	8	10
2	5	11
3	9	13
4	7	15
5	11	17
6	13	20
7	21	21
8	16	24
9	21	26
10	22	29

Задание

1. Изобразить диаграмму рассеяния и сформулировать гипотезу о форме связи.
2. Найти параметры a и b уравнения парной линейной регрессии $y = a + bx$. Пояснить эконометрический смысл параметра b.
3. Оценить статистическую значимость коэффициента регрессии (b) используя t-критерий Стьюдента на уровне значимости $\alpha = 0,05$.
4. Рассчитать границы доверительного интервала для параметра b.
5. Вычислить коэффициент корреляции r и оценить тесноту связи между фактором и откликом.
6. Вычислить коэффициент детерминации R^2, пояснить его эконометрический смысл и проверить его значимость с использованием F-критерия Фишера при $\alpha = 0,05$.
7. Проанализировать остатки.
8. Вычислить MAPE.
9. Выяснить, в какой интервал с вероятностью 95% попадет прогнозное значение отклика, если значение фактора равно $\bar{x}$.

<u>Решение:</u>

1. **Графический анализ** – построение диаграммы рассеяния, по которой определяется форма регрессионной модели.



По расположению точек предположим наличие линейной зависимости $y = a + bx$.

2. **Вычислим параметры** a и b уравнения парной линейной регрессии $y = a + bx$.

Для расчета параметров уравнения линейной регрессии составляем расчетную таблицу. Сначала заполняем столбцы с (1) по (5).

№	y	x	xy	x^2	y^2	y	$e = y - y$	e^2	$\left\lvert\frac{e}{y}\right\rvert \cdot 100\%$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	8	10	80	100	64	6	2	4	25,00
2	5	11	55	121	25	7	-2	4	40,00
3	9	13	117	169	81	8	1	1	11,11
4	7	15	105	225	49	10	-3	9	42,86
5	11	17	187	289	121	12	-1	1	9,09
6	13	20	260	400	169	15	-2	4	15,38
7	21	21	441	441	441	16	5	25	23,81
8	16	24	384	576	256	18	-2	4	12,50
9	21	26	546	676	441	20	1	1	4,76
10	22	29	638	841	484	23	-1	1	4,55
Сумма	133	186	2813	3838	2131		-2	54	189,06
Среднее	13,3	18,6	281,3	383,8	213,1				18,91

$$b = \frac{\overline{yx} - \bar{y} \cdot \bar{x}}{\overline{x^2} - \left(\bar{x}\right)^2} = \frac{281,3 - 18,6 \cdot 13,3}{383,3 - 18,6^2} = 0,9;$$

$$a = \bar{y} - b \cdot \bar{x} = 13,3 - 0,9 \cdot 18,6 = -3,4.$$

Получено уравнение регрессии: $y = -3,4 + 0,9x$.

Эконометрический смысл коэффициента регрессии: с увеличением времени выбора товара на 1 минуту количество купленных товаров возрастает в среднем на 0,9 шт. (При интерпретации можно 0,9 шт. округлить до 1, т.е. каждая проведенная в магазине минута прибавляет к покупке в среднем 1 единицу товара.)

3. **Проверим статистическую значимость** коэффициента регрессии.

Используем *t*-критерий Стьюдента. Выдвигаем гипотезу $H_0: b = 0$ об отсутствии влияния фактора на отклик. Далее, необходимо сначала заполнить столбцы с (6) по (8), затем вычислить стандартную ошибку:

$$SE_b = \sqrt{\frac{\sum e_i^2}{(N-1)\cdot(N-2)\cdot\sigma_x^2}} = \left|\sigma_x = \sqrt{\frac{N}{N-1}\cdot\left(\overline{x^2}-\left(\overline{x}\right)^2\right)} = \sqrt{\frac{10}{10-1}\cdot\left(393{,}3-13{,}3^2\right)} = 6{,}44\right| =$$

$$= \sqrt{\frac{54}{(10-1)\cdot(10-2)\cdot 6{,}44^2}} = 0{,}134.$$

Фактическое значение *t*-критерия Стьюдента: $t_b = \frac{|b|}{SE_b} = \frac{0{,}9}{0{,}134} = 6{,}716$.

$t_{табл}$ на уровне значимости $\alpha = 0{,}05$ и числа степеней свободы $N-2 = 10-2 = 8$ равно 2,306.

$t_b = 6{,}716 > t_{табл} = 2{,}306$, гипотеза H_0 отклоняется, т. е. влияние фактора на отклик обнаружено.

4. **Границы 95-процентного доверительного интервала** для коэффициента регрессии:

$н.гр. = b - t_{табл}\cdot SE_b = 0{,}9 - 2{,}306\cdot 0{,}134 = 0{,}519,$

$в.гр. = b + t_{табл}\cdot SE_b = 0{,}9 + 2{,}306\cdot 0{,}134 = 1{,}209.$

При увеличении времени выбора товара на 1 мин. Количество выбранных товаров вырастет в среднем на 0,9 шт., в 95% случаев рост может составлять от 0,519 шт. до 1,209 шт. (В данном случае границы доверительного интервала округляются до 1 шт. в соответствии со смыслом отклика.)

5. **Вычислим коэффициент корреляции**:

$$r = b\frac{\sigma_x}{\sigma_y} = \left|\sigma_y = \sqrt{\frac{N}{N-1}\cdot\left(\overline{y^2}-\left(\overline{y}\right)^2\right)} = \sqrt{\frac{10}{10-1}\cdot\left(213{,}1-13{,}3^2\right)} = 6{,}34\right| =$$

$$= 0{,}9\cdot\frac{6{,}44}{6{,}34} = 0{,}914.$$

Корреляция больше нуля, значит связь прямая, по шкале Чеддока – весьма высокая.

6. **Вычислим коэффициент детерминации**: $R^2 = 0{,}914^2 = 0{,}835$. Это означает, что 67,7% количества приобретенных товаров объясняется временем выбора товара. $R^2 = 83{,}5\% > 30\%$, значит прогнозировать по данной модели целесообразно.

Проверим статистическую значимость уравнения регрессии с помощью *F*-критерия Фишера. Выдвигаем гипотезу H_0 о статистической незначимости уравнения регрессии и коэффициента детерминации. Фактическое значение *F*-критерия равно:

$$F_{факт} = \frac{R^2}{1-R^2}(N-2) = \frac{0{,}835}{1-0{,}835}(10-2) = 40{,}49.$$

$F_{табл} = 5{,}318$ на уровне значимости $\alpha = 0{,}05$ и числе степеней свободы 1 и

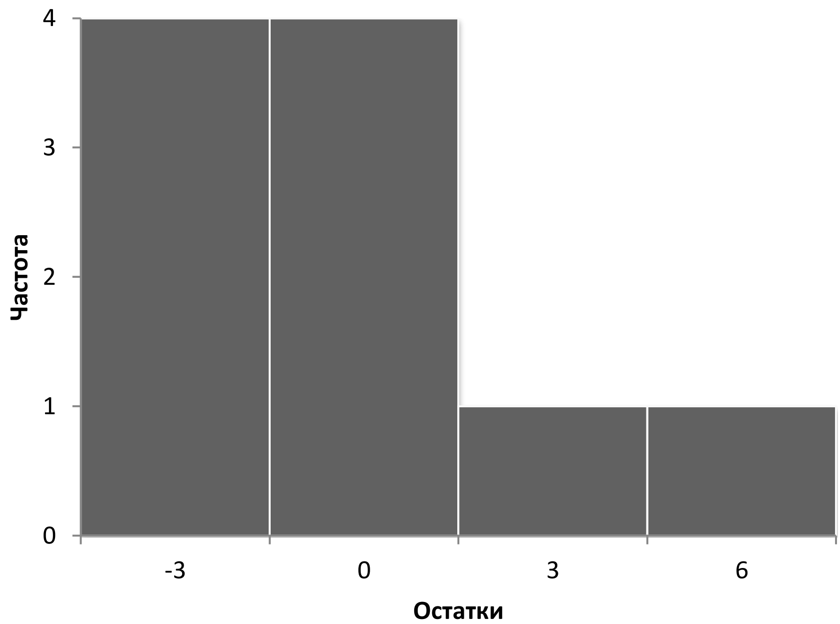
$N-2=10-2=8$.

$F_{факт}=40,49>F_{табл}=5,318$, гипотеза H_0 отклоняется и признается статистическая значимость уравнения регрессии. Построенная модель «лучше» прогноза по среднему.

7. **Проанализируем остатки**.
Проверим остатки на нормальность графическим способом:
1) Изобразим гистограмму. Для этого построим интервальный вариационный ряд. Число интервалов, на которые разобьем найденные остатки, определим по формуле Стерджесса
$n=1+3,322\cdot\lg N=1+3,322\cdot\lg 10\approx 4$.
Длина интервала: $h=\frac{R}{n-1}=\frac{e_{\max}-e_{\min}}{n-1}=\frac{5-(-3)}{4-1}=2,67\approx 3$.
Границы первого интервала определим следующим образом:
$\left[e_{\min}-\frac{h}{2};e_{\min}+\frac{h}{2}\right]=\left[-3-\frac{3}{2};-3+\frac{3}{2}\right]=\left[-4,5;-1,5\right]$.

№	Нижняя граница	Верхняя граница	Частота
1	-4,5	-1,5	4
2	-1,5	1,5	4
3	1,5	4,5	1
4	4,5	7,5	1
Сумма			10

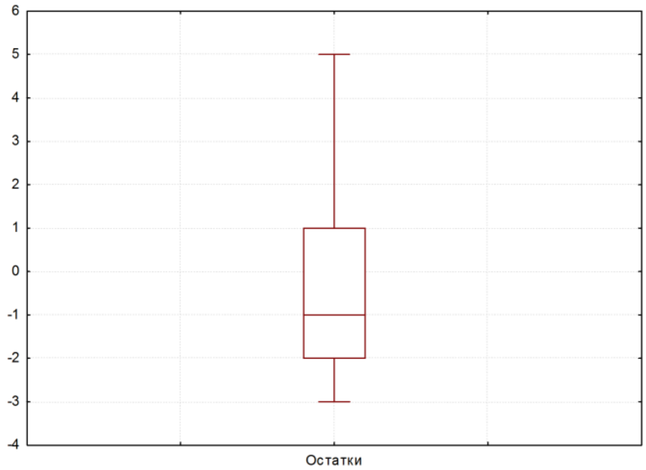
Гистограмма будет иметь вид:



2) Изобразим ящик с усами. Для этого упорядочим остатки по возрастанию и найдем медиану и квартили.

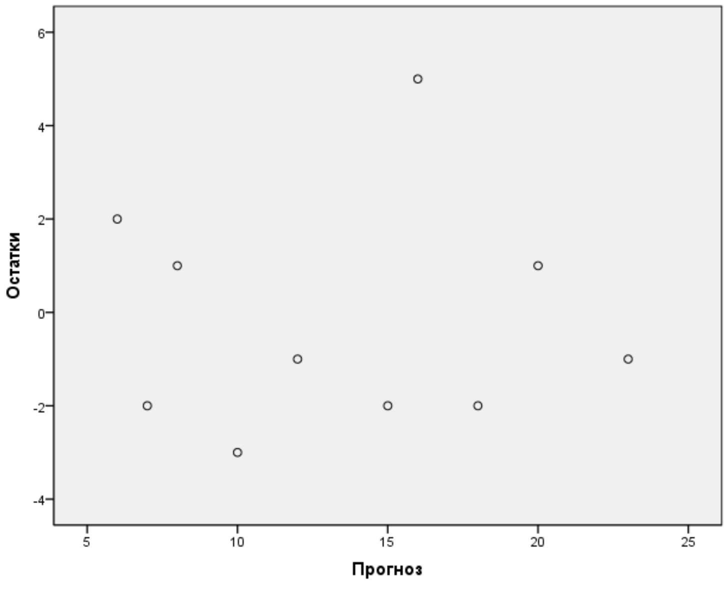
Остатки	Упорядоченные остатки	
2	-3	минимум
-2	-2	
1	-2	$=Q_н$
-3	-2	
-1	-1	$Me=\dfrac{-1+(-1)}{2}=-1$
-2	-1	
5	1	
-2	1	$=Q_в$
1	2	
-1	5	максимум

Ящик с усами будет иметь вид:



С учетом недостаточного размера выборки будем считать гистограмму и ящик с усами симметричными.

С помощью диаграммы рассеяния проверим, чтобы остатки не зависели от предсказанных по уравнению регрессии значений.



Остатки не зависят от предсказанных по уравнению регрессии значений.

Все это говорит о приемлемости линейной модели для описания скрытой в исходных данных зависимости.

8. **Оценим точность прогнозов**, вычислим среднюю абсолютную ошибку прогноза в процентах (MAPE). Заполнив столбец (9) в расчетной таблице, получим $MAPE = 18,91\%$. Это означает, что при прогнозе по построенной модели ошибка в среднем будет составлять 18,91%. Такой результат будем считать приемлемым.
9. **Вычислим прогнозное значение отклика** y_f, если прогнозное значение фактора $x_f = \bar{x} = 18,6$ мин.

   $y_f = -3,4 + 0,9 \cdot 18,6 = 13,34 \cong 13$ шт.

   Стандартная ошибка прогноза:

$$SE_f = \sqrt{\frac{\sum e^2}{N-2} \cdot \left(1 + \frac{1}{N} + \frac{(x_f - \bar{x})^2}{\sigma_x^2 \cdot (N-1)}\right)} = \sqrt{\frac{54}{10-2} \cdot \left(1 + \frac{1}{10} + \frac{(18,6-18,6)^2}{6,44^2 \cdot (10-1)}\right)} = 2,725.$$

   95-процентный доверительный интервал прогноза строится по формуле:

$$н.гр. = y_f - t_{табл} \cdot SE_f = 13 - 2,306 \cdot 2,725 = 6,716 \cong 7,$$

$$в.гр. = y_f + t_{табл} \cdot SE_f = 13 + 2,306 \cdot 2,725 = 19,284 \cong 19.$$

   Для времени выбора товара, равного 18,6 мин. количество приобретенных товаров с вероятностью 0,95 будет составлять от 7 шт. до 19 шт.