Пример выполнения контрольной работы

Постановка задачи

Имеются следующие данные о покупателях продуктовых магазинов некоторой сети:

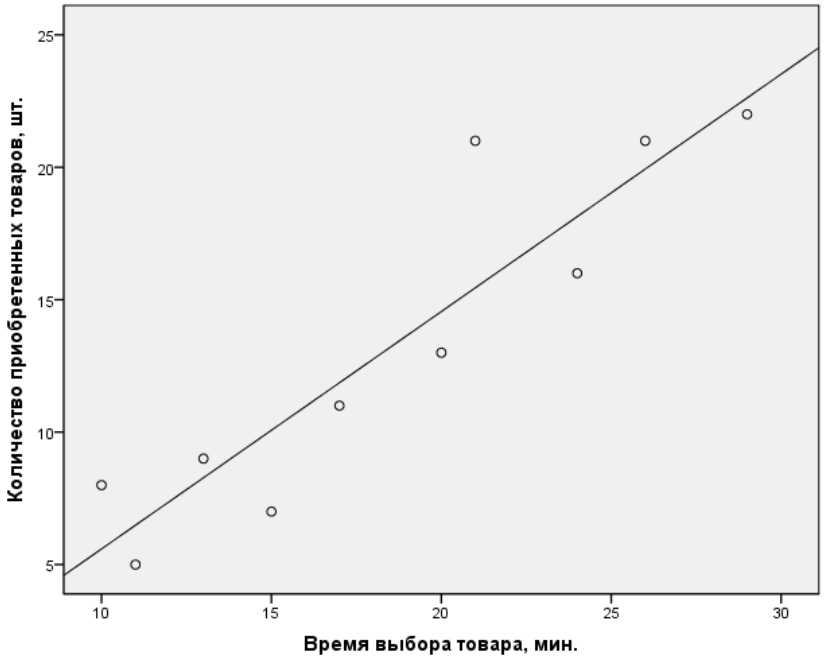
№ покупателя	Количество приобретенных товаров, шт., y	Время выбора товара, мин., x
1	8	10
2	5	11
3	9	13
4	7	15
5	11	17
6	13	20
7	21	21
8	16	24
9	21	26
10	22	29

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 $\bar{x}$.

Решение:

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	$\left\|\frac{e}{y}\right\| \cdot 100\%$
	(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

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

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

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

$n=1+3,322\cdot\lg N=1+3,322\cdot\lg 10\approx 4$.

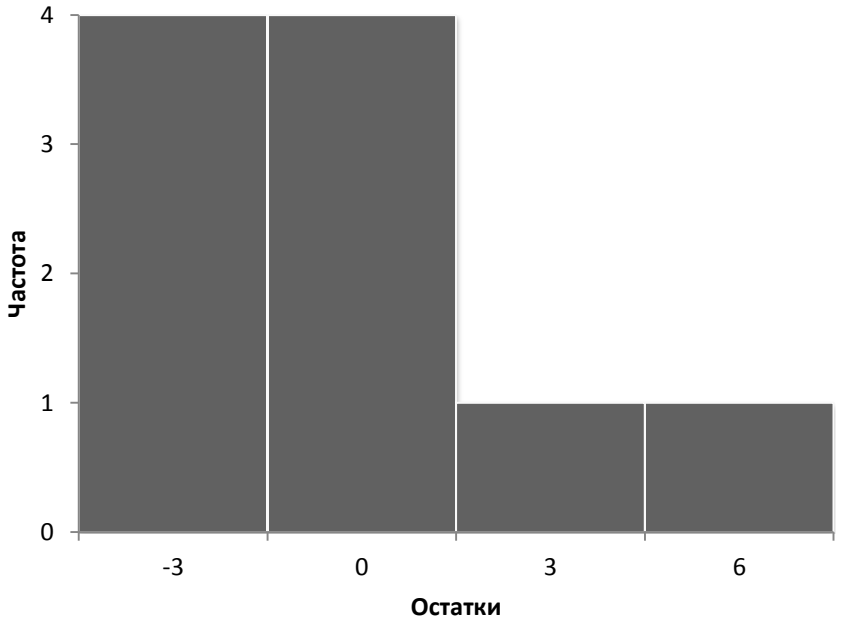
Длина интервала: $h=\dfrac{R}{n-1}=\dfrac{e_{\max}-e_{\min}}{n-1}=\dfrac{5-(-3)}{4-1}=2,67\approx 3$.

Границы первого интервала определим следующим образом:

$$\left[e_{\min}-\frac{h}{2};e_{\min}+\frac{h}{2}\right]=\left[-3-\frac{3}{2};-3+\frac{3}{2}\right]=[-4,5;-1,5].$$

№	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

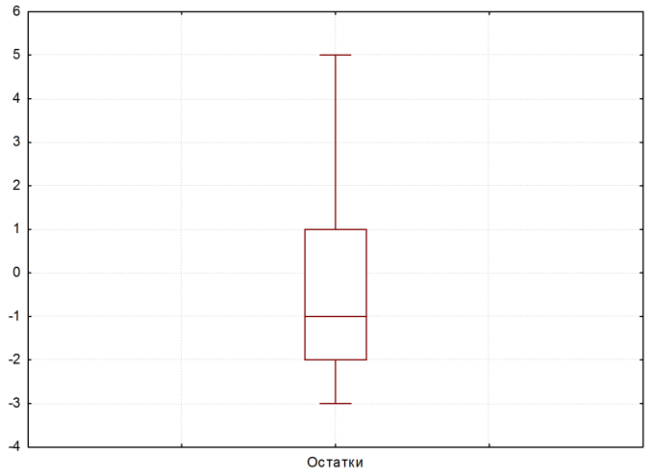
Гистограмма будет иметь вид:



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

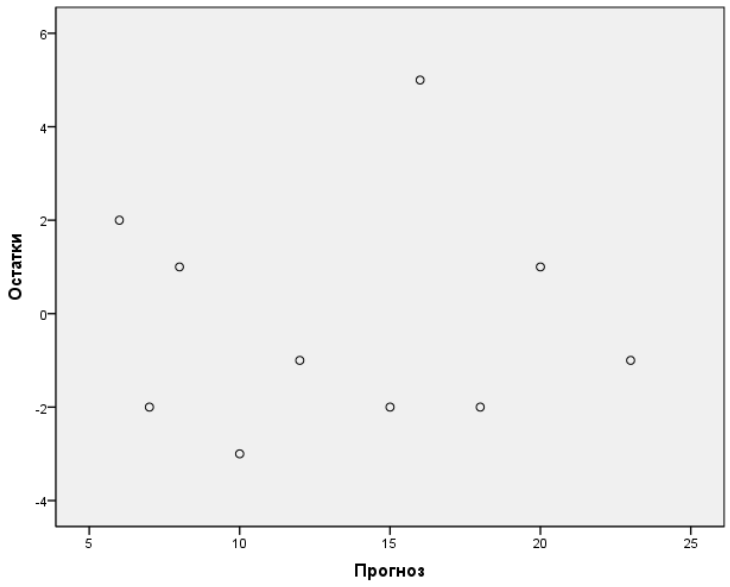
Остатки	Упорядоченные остатки	
2	-3	минимум
-2	-2	
1	-2	$= Q_н$
-3	-2	
-1	-1	$Me = \frac{-1+(-1)}{2} = -1$
-2	-1	
5	1	
-2	1	$= Q_в$
1	2	
-1	5	максимум

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.