РЕШЕНИЕ ЗАДАЧ НА ПОСТРОЕНИЕ ЛИНЕЙНОЙ РЕГРЕССИИ

ЗАДАНИЕ.
В таблице 2 приведены данные зависимости потребления Y (усл. ед.) от дохода X (усл. ед.) для некоторых домашних хозяйств.

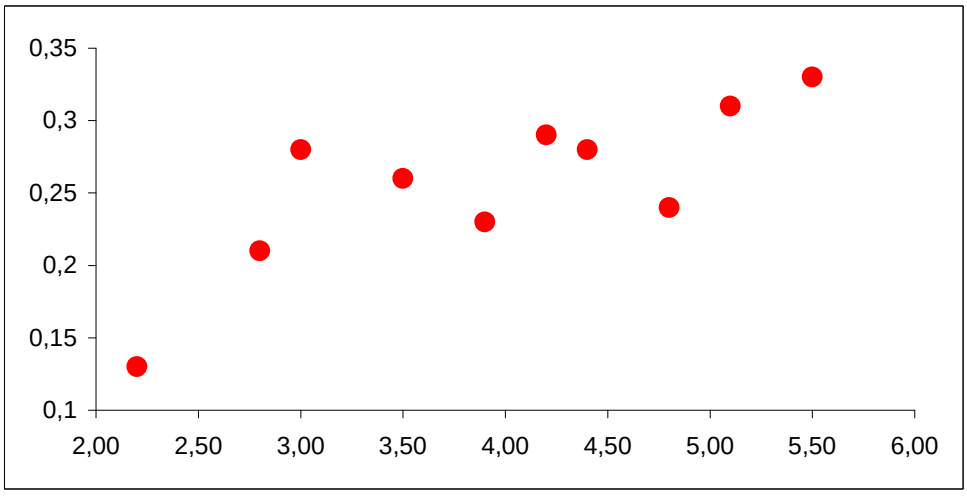
1. В предположении, что между Y и X существует линейная зависимость, найдите точечные оценки коэффициентов линейной регрессии.
2. Найдите стандартное отклонение s и коэффициент детерминации R^2.
3. В предположении нормальности случайной составляющей регрессионной модели проверьте гипотезу об отсутствии линейной зависимости между Y и X.
4. Каково ожидаемое потребление $\widehat{Y}_n$ домашнего хозяйства с доходом $X_n = 7$ усл. ед.? Найдите доверительный интервал для прогноза.

Дайте интерпретацию полученных результатов. Уровень значимости во всех случаях считать равным $\alpha = 1 - \gamma = 0,05$.

X	2,20	2,80	3,00	3,50	3,90	4,20	4,40	4,80	5,10	5,50
Y	0,13	0,21	0,28	0,26	0,23	0,29	0,28	0,24	0,31	0,33

РЕШЕНИЕ.

Построим облако корреляции, точки на плоскости, где абсцисса будет означать доход, а ордината – объем потребления. Получим:



По виду диаграммы можно предположить, что между доходами и объемом потребления существует прямая линейная связь.

Сначала найдем характеристики случайных величин X и Y: выборочное среднее и выборочное среднее квадратическое отклонение.

Выборочная средняя $\bar{x} = \frac{1}{n}\sum x_i = \frac{1}{10}39,40 = 3,94$

Выборочная дисперсия $\overline{D}_x = \frac{1}{n}\sum (x_i - \bar{x})^2 = \frac{1}{10}10,2 = 1,02$

Выборочное квадратическое отклонение $\sigma_x = \sqrt{\overline{D_x}} = 1,01$

Выборочная средняя $\bar{y} = \frac{1}{n}\sum y_i = \frac{1}{10}2,56 = 0,256$

Выборочная дисперсия $\overline{D}_y = \frac{1}{n}\sum (y_i - \bar{y})^2 = \frac{1}{10}0,03 = 0,003$

Выборочное квадратическое отклонение $\sigma_y = \sqrt{\overline{D_y}} = 0,054$

Осталось подсчитать $\sum x_i y_i = 10,52$.

Все расчеты занесем в таблицу:

											Сумма
x_i	2,20	2,80	3,00	3,50	3,90	4,20	4,40	4,80	5,10	5,50	**39,40**
y_i	0,13	0,21	0,28	0,26	0,23	0,29	0,28	0,24	0,31	0,33	**2,56**
$(x_i - \bar{x})^2$	3,0276	1,2996	0,8836	0,1936	0,0016	0,0676	0,2116	0,7396	1,3456	2,4336	**10,20**
$(y_i - \bar{y})^2$	0,0159	0,0021	0,0006	2E-05	0,0007	0,0012	0,0006	0,0003	0,0029	0,0055	**0,03**
$x_i y_i$	0,286	0,588	0,84	0,91	0,897	1,218	1,232	1,152	1,581	1,815	**10,52**

Тогда линейный коэффициент корреляции
$r_в = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{n\sigma_x\sigma_y} = \frac{10,52 - 10 \cdot 3,94 \cdot 0,256}{10 \cdot 1,01 \cdot 0,054} \approx 0,787$. Связь тесная, прямая.

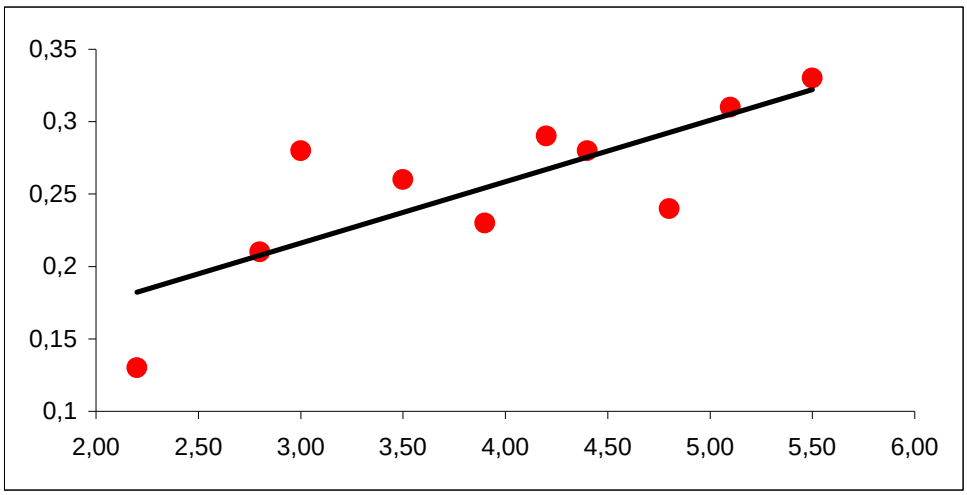
Уравнение регрессии Y на X имеет вид $\overline{y_x} - \bar{y} = r_в \frac{\sigma_y}{\sigma_x}(x - \bar{x})$. Подставляем все величины:

$\overline{y_x} - 0,256 = 0,787\frac{0,054}{1,01}(x - 3,94)$

$\overline{y_x} = 0,042x + 0,09$

Точечные оценки параметров регрессии: $a = 0,042$, $b = 0,09$.

Графическое изображение поля корреляции и линии регрессии:



Найдем стандартное отклонение s.

$$s_x = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2} = \sqrt{\frac{1}{9}10,2} \approx 1,065,$$

$$s_y = \sqrt{\frac{1}{n-1}\sum (y_i - \bar{y})^2} = \sqrt{\frac{1}{9}0,03} \approx 0,057$$

Коэффициент детерминации $R^2 = r_в^2 = 0,787^2 \approx 0,619$.

Введем нулевую гипотезу $H_0 : r = 0$. Проверим гипотезу об отсутствии линейной корреляционной зависимости (о незначимости коэффициента корреляции). Вычислим значение критерия $T_{набл} = \frac{r_в\sqrt{n-2}}{\sqrt{1-r_в^2}} = \frac{0,787\cdot\sqrt{9}}{\sqrt{1-(0,787)^2}} \approx 3,83$.

Найдем критическую точку по уровню значимости $\alpha = 0,05$ и числу степеней свободы $k = n-2 = 8$, получаем $t_{кр} = 2,306$. Так как $|T_{набл}| = 3,83 > 2,306 = t_{кр}$, следует отвергнуть нулевую гипотезу $H_0 : r = 0$, то есть существует линейная зависимость между доходами и объемом потребления, коэффициент корреляции статистически значим.

Найдем, каково ожидаемое потребление $\widehat{Y_n}$ домашнего хозяйства с доходом $X_n = 7$ усл. ед.: $\widehat{Y_n}(7) = 0,042\cdot 7 + 0,09 = 0,384$ усл. ед. – точечная оценка.

Найдем доверительный интервал для прогноза. Вычислим среднюю стандартную ошибку прогноза по формуле:

$$m_{\hat{y}_n} = \sigma_{ост}\sqrt{1+\frac{1}{n}+\frac{(x_n-\bar{x})^2}{n\sum_{i=1}^{n}(x-\bar{x})^2}} = \sqrt{\frac{\sum_{i=1}^{n}(y-\hat{y})^2}{n-2}}\cdot\sqrt{1+\frac{1}{n}+\frac{(x_n-\bar{x})^2}{n\sigma_x^2}}$$

Здесь $n=10$, $x_n=7$, $\bar{x}=3,94$.

$$\sum_{i=1}^{n}(y-\hat{y})^2=\sum_{i=1}^{10}(y-\hat{y})^2=0,011,$$

$$\sum_{i=1}^{n}(x-\bar{x})^2=\sum_{i=1}^{12}(x-\bar{x})^2=10,2.$$

Дополнительные расчеты ниже:

											Сумма
x_i	2,20	2,80	3,00	3,50	3,90	4,20	4,40	4,80	5,10	5,50	**39,400**
y_i	0,13	0,21	0,28	0,26	0,23	0,29	0,28	0,24	0,31	0,33	**2,560**
$\hat{y}$	0,182	0,208	0,216	0,237	0,254	0,266	0,275	0,292	0,304	0,321	**2,555**
$(y-\hat{y})^2$	0,003	6E-06	0,004	5E-04	6E-04	6E-04	3E-05	0,003	3E-05	8E-05	**0,011**

Получаем:

$$m_{\hat{y}_n}=\sqrt{\frac{0,011}{8}}\cdot\sqrt{1+\frac{1}{10}+\frac{(7-3,94)^2}{10\cdot 10,2}}\approx 0,0405.$$

Определяем по таблице Стьюдента t по параметрам: $k=n-2=8$ степеней свободы, $\alpha=0,05$, откуда $t=2,31$

Получаем интервал для прогнозного значения:

$$\widehat{y_n}-t\cdot m_{\hat{y}_n}<y_n<\widehat{y_n}+t\cdot m_{\hat{y}_n},$$

$$0,384-2,31\cdot 0,0405<y_n<0,384+2,31\cdot 0,0405,$$

$$0,29<y_n<0,478$$

Дадим интерпретацию полученных результатов.
По результатам исследования можно сделать вывод, что между доходами и объемом потребления существует прямая линейная связь, которая является достаточно тесной (коэффициент корреляции 0,787). Доля дисперсии признака Y в общей дисперсии Y, объясненную регрессией Y по X, которая выражается через коэффициент детерминации 0,619 говорит о том, что линейная модель адекватна статистике на 61,9%.