

Варианты индивидуальных заданий

D.1. Парная регрессия и корреляция

Пример. По территориям региона приводятся данные за 199X г.

Таблица D.1

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного, руб., x	Среднедневная заработная плата, руб., y
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

Требуется:

1. Построить линейное уравнение парной регрессии y по x .
2. Рассчитать линейный коэффициент парной корреляции, коэффициент детерминации и среднюю ошибку аппроксимации.
3. Оценить статистическую значимость уравнения регрессии в целом и отдельных параметров регрессии и корреляции с помощью F -критерия Фишера и t -критерия Стьюдента.
4. Выполнить прогноз заработной платы y при прогнозном значении среднедушевого прожиточного минимума x , составляющем 107% от среднего уровня.
5. Оценить точность прогноза, рассчитав ошибку прогноза и его доверительный интервал.

6. На одном графике отложить исходные данные и теоретическую прямую.

Решение

1. Для расчета параметров уравнения линейной регрессии строим расчетную таблицу D.2.

Таблица D.2

№	x	y	$y \cdot x$	x^2	y^2	\hat{y}_x	$y - \hat{y}_x$	$(y - \hat{y}_x)^2$	A_i
1	2	3	4	5	6	7	8	9	10
1	78	133	10374	6084	17689	148,78	-15,78	249,01	11,86
2	82	148	12136	6724	21904	152,46	-4,46	19,89	3,01
3	87	134	11658	7569	17956	157,06	-23,06	531,76	17,21
4	79	154	12166	6241	23716	149,70	4,30	18,49	2,79
5	89	162	14418	7921	26244	158,90	3,10	9,61	1,91
6	106	195	20670	11236	38025	174,54	20,46	418,61	10,49
7	67	139	9313	4489	19321	138,66	0,34	0,12	0,24
8	88	158	13904	7744	24964	157,98	0,02	0,00	0,01
9	73	152	11096	5329	23104	144,18	7,82	61,15	5,14
10	87	162	14094	7569	26244	157,06	4,94	24,40	3,05
11	76	159	12084	5776	25281	146,94	12,06	145,44	7,58
12	115	173	19895	13225	29929	182,82	-9,82	96,43	5,68
Итого	1027	1869	161808	89907	294377	1869,08	-0,08	1574,91	68,97
Среднее значение	85,58	155,75	13484,0	7492,25	24531,4	155,76	—	131,24	5,75
σ	12,97	16,53	—	—	—	—	—	—	—
σ^2	168,31	273,34	—	—	—	—	—	—	—

Находим оценки параметров уравнения регрессии:

$$b = \frac{\overline{y \cdot x} - \bar{y} \cdot \bar{x}}{\overline{x^2} - \bar{x}^2} = \frac{13484 - 155,75 \cdot 85,58}{7492,25 - 85,58^2} = \frac{154,915}{168,31} = 0,92;$$

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

Получено уравнение регрессии: $\hat{y} = 77,02 + 0,92 \cdot x$.

С увеличением среднедушевого прожиточного минимума на 1 руб. среднедневная заработная плата возрастает в среднем на 0,92 руб.

2. Тесноту линейной связи оценит коэффициент корреляции:

$$r_{xy} = b \cdot \frac{\sigma_x}{\sigma_y} = 0,92 \cdot \frac{12,97}{16,53} = 0,722;$$

Т.к. значение коэффициента корреляции больше 0,7, то это говорит о наличии весьма тесной линейной связи между признаками.

Коэффициент детерминации:

$$r_{xy}^2 = 0,521.$$

Это означает, что 52% вариации заработной платы (y) объясняется вариацией фактора x – среднедушевого прожиточного минимума.

Качество модели определяет средняя ошибка аппроксимации (2,7):

$$\bar{A} = \frac{1}{n} \sum A_i = \frac{68,97}{12} = 5,75\%.$$

Качество построенной модели оценивается как хорошее, так как \bar{A} не превышает 10%.

3. Оценку значимости уравнения регрессии в целом проведем с помощью F -критерия Фишера. Фактическое значение F -критерия:

$$F_{\text{факт}} = \frac{r_{xy}^2}{1 - r_{xy}^2} \cdot (n - 2) = \frac{0,521}{1 - 0,521} \cdot 10 = 10,88.$$

Табличное значение критерия при пятипроцентном уровне значимости и степенях свободы $k_1 = 1$ и $k_2 = 12 - 2 = 10$ составляет $F_{\text{табл}} = 4,96$. Так как $F_{\text{факт}} = 10,88 > F_{\text{табл}} = 4,96$, то уравнение регрессии признается статистически значимым.

Оценку статистической значимости параметров регрессии проведем с помощью t -статистики Стьюдента и путем расчета доверительного интервала каждого из показателей.

Табличное значение t -критерия для числа степеней свободы $df = n - 2 = 12 - 2 = 10$ и $\alpha = 0,05$ составит $t_{\text{табл}} = 2,23$.

Определим случайные ошибки m_a , m_b , $m_{r_{xy}}$ (остаточная дисперсия

на одну степень свободы $S_{\text{ост}}^2 = \frac{\sum (y - \hat{y}_x)^2}{n - 2} = \frac{1574,91}{10} = 157,49$):

$$m_a = \sqrt{S_{\text{ост}}^2 \frac{\sum x^2}{n^2 \sigma_x^2}} = \sqrt{157,49 \cdot \frac{89907}{12^2 \cdot 164,94}} = 24,42;$$

$$m_b = \sqrt{\frac{S_{\text{ост}}^2}{n \cdot \sigma_x^2}} = \sqrt{\frac{157,49}{12 \cdot 164,94}} = 0,282;$$

$$m_{r_{xy}} = \sqrt{\frac{1 - r_{xy}^2}{n - 2}} = \sqrt{\frac{1 - 0,521}{12 - 2}} = 0,219.$$

Тогда

$$t_a = \frac{a}{m_a} = \frac{77,02}{24,42} = 3,15;$$

$$t_b = \frac{b}{m_b} = \frac{0,92}{0,282} = 3,26;$$

$$t_{r_{xy}} = \frac{r_{xy}}{m_{r_{xy}}} = \frac{0,722}{0,219} = 3,30.$$

Фактические значения t -статистики превосходят табличное значение:

$$t_a = 3,26 > t_{\text{табл}} = 2,3; \quad t_b = 3,16 > t_{\text{табл}} = 2,3; \quad t_{r_{xy}} = 3,25 > t_{\text{табл}} = 2,3,$$

поэтому параметры a , b и r_{xy} не случайно отличаются от нуля, а статистически значимы.

Рассчитаем доверительные интервалы для параметров регрессии a и b . Для этого определим предельную ошибку для каждого показателя:

$$\Delta_a = t_{\text{табл}} \cdot m_a = 2,23 \cdot 24,42 = 54,46;$$

$$\Delta_b = t_{\text{табл}} \cdot m_b = 2,23 \cdot 0,282 = 0,63.$$

Доверительные интервалы

$$\gamma_a = a \pm \Delta_a = 77,02 \pm 54,46 \text{ и } 22,56 \leq a^* \leq 131,48;$$

$$\gamma_b = b \pm \Delta_b = 0,92 \pm 0,63 \text{ и } 0,29 \leq b^* \leq 1,55$$

Анализ верхней и нижней границ доверительных интервалов приводит к выводу о том, что с вероятностью $p = 1 - \alpha = 0,95$ параметры a и b , находясь в указанных границах, не принимают нулевых значений, т.е. являются статистически значимыми и существенно отличны от нуля.

4. Полученные оценки уравнения регрессии позволяют использовать его для прогноза. Если прогнозное значение прожиточного минимума составит: $x_p = \bar{x} \cdot 1,07 = 85,6 \cdot 1,07 = 91,6$ руб., тогда индивидуальное прогнозное значение заработной платы составит: $\hat{y}_0 = 77,02 + 0,92 \cdot 91,6 = 161,29$ руб.

5. Ошибка прогноза составит:

$$m_{\hat{y}_0} = \sqrt{S_{\text{ост}}^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{n \cdot \sigma_x^2} \right)} = \sqrt{157,49 \cdot \left(1 + \frac{1}{12} + \frac{(91,6 - 85,6)^2}{12 \cdot 164,94} \right)} = 13,17.$$

Предельная ошибка прогноза, которая в 95% случаев не будет превышена, составит:

$$\Delta_{\hat{y}_0} = t_{\text{табл}} \cdot m_{\hat{y}_0} = 2,23 \cdot 13,17 = 29,37.$$

Доверительный интервал прогноза:

$$\gamma_{\hat{y}_0} = \hat{y}_0 \pm \Delta_{\hat{y}_0} = 161,29 \pm 29,37 \text{ и } 131,92 \leq y_0^* \leq 190,66.$$

Выполненный прогноз среднемесячной заработной платы является надежным ($p = 1 - \alpha = 1 - 0,05 = 0,95$) и находится в пределах от 131,92 руб. до 190,66 руб.

6. В заключение решения задачи построим на одном графике исходные данные и теоретическую прямую (рис. D.1):

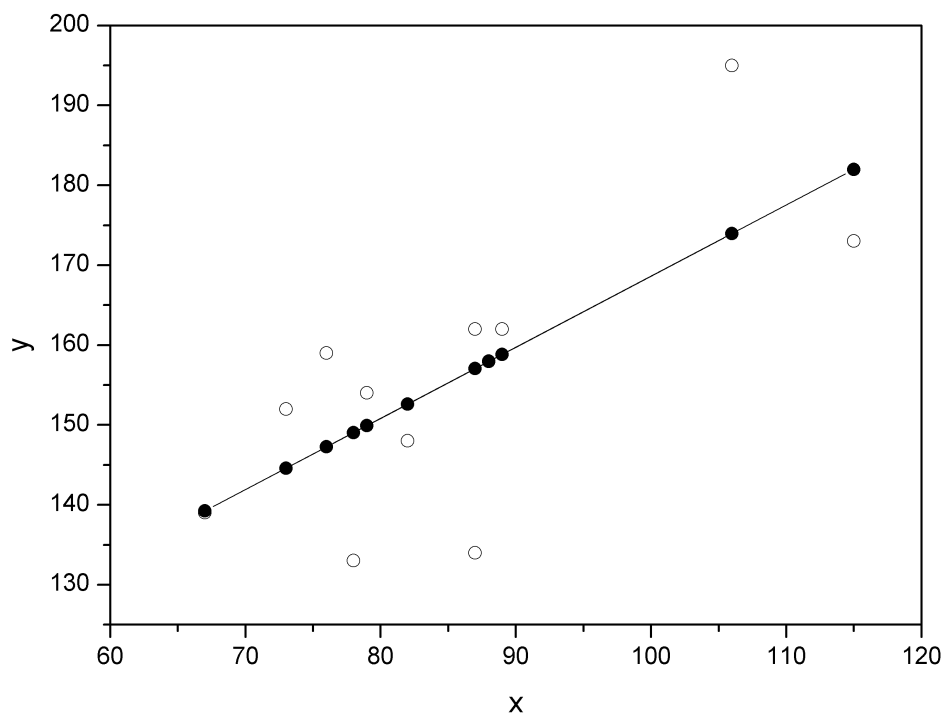


Рис. D.1.

Варианты индивидуальных заданий

Задача 1. По территориям региона приводятся данные за 199X г. (см. таблицу своего варианта).

Требуется:

1. Построить линейное уравнение парной регрессии y от x .
2. Рассчитать линейный коэффициент парной корреляции и среднюю ошибку аппроксимации.
3. Оценить статистическую значимость параметров регрессии и корреляции с помощью F -критерия Фишера и t -критерия Стьюдента.
4. Выполнить прогноз заработной платы y при прогнозном значении среднедушевого прожиточного минимума x , составляющем 107% от среднего уровня.
5. Оценить точность прогноза, рассчитав ошибку прогноза и его доверительный интервал.

6. На одном графике построить исходные данные и теоретическую прямую.

Вариант 1

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного, руб., x	Среднедневная заработная плата, руб., y
1	81	124
2	77	131
3	85	146
4	79	139
5	93	143
6	100	159
7	72	135
8	90	152
9	71	127
10	89	154
11	82	127
12	111	162

Вариант 2

Номер региона	Среднедушевой прожиточный минимум в день одного трудоспособного, руб., x	Среднедневная заработная плата, руб., y
1	74	122
2	81	134
3	90	136
4	79	125
5	89	120
6	87	127
7	77	125
8	93	148
9	70	122
10	93	157
11	87	144
12	121	165