

Продолжение лекции:

Наблюдаемые значения объясняемой переменной y_i ($i=\overline{1,n}$) отличаются от прогнозируемых значений $\hat{y}_i$ $i=\overline{1,n}$, рассчитанных по уравнению регрессии (2.3). Чем меньше эти отличия, тем точнее прогнозируемые значения $\hat{y}_i$ описывают наблюдаемые значения y_i , и тем лучше качество построенной модели. Величина отклонения наблюдаемого и прогнозируемого значений объясняемой переменной $(y_i - \hat{y}_i)$ по каждому наблюдению представляет собой ошибку аппроксимации. Так как $(y_i - \hat{y}_i)$ может быть как величиной положительной, так и отрицательной, то ошибку аппроксимации для каждого наблюдения принято определять в процентах и по модулю.

Выражение $(y_i - \hat{y}_i)$ можно рассматривать как абсолютную ошибку аппроксимации, а выражение:

$$A_i = \left| \frac{(y_i - \hat{y}_i)}{y_i} \right| \cdot 100,$$

как относительную ошибку аппроксимации для i -го наблюдения.

Чтобы иметь показатель, характеризующий качество модели в целом, определяется средняя относительная ошибка аппроксимации по всем наблюдениям в выборке по формуле:

$$\bar{A} = \frac{1}{n} \cdot \sum_{i=1}^n A_i = \frac{1}{n} \cdot \sum_{i=1}^n \left| \frac{(y_i - \hat{y}_i)}{y_i} \right| \cdot 100.$$

Считается [3], что построенное уравнение регрессии достаточно хорошо прогнозирует наблюдаемые значения объясняемой переменной, если $\bar{A} < (10 \div 12)\%$ и удовлетворительно, если $\bar{A} < (12 \div 15)\%$.

Пример решения задачи модели парной линейной регрессии.

Пример 2.1. Известны данные о сменной добычей торфа на одного рабочего $Y(m)$ и мощности пласта $X(m)$ по следующим (условным) данным (таблица 2.1), характеризующим процесс добычи торфа в $n = 16$ карьерах. Необходимо построить уравнение парной регрессии, описывающее зависимость массы добываемого торфа в карьере на одного рабочего в смену от толщины пласта торфа.

Таблица 2.1.

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
x_i	4	7	5	8	6	7	9	6	5	7	9	8	7	8	7	9
y_i	11	13	12	13	15	15	17	14	12	15	17	16	15	17	19	18

Для определения вида зависимости между X и Y построим корреляционное поле (смотрите рис.2.2):

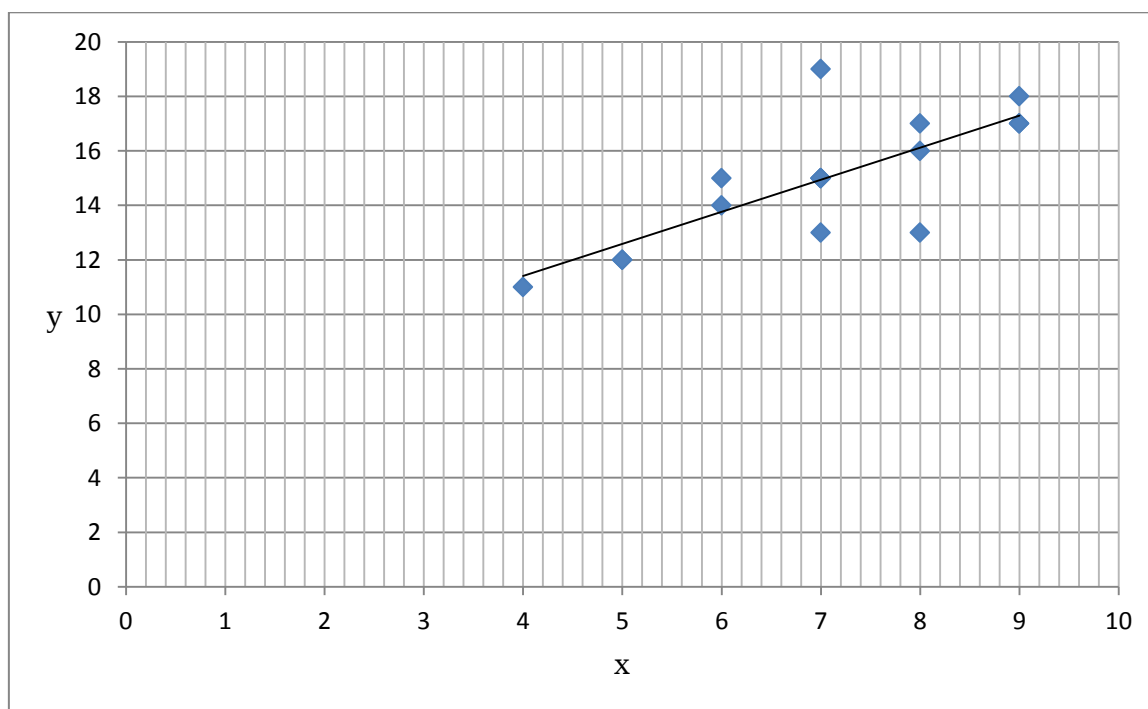


Рис. 2.2 – Корреляционное поле.

По расположению точек на корреляционном поле можно сделать предположение, что зависимость между X и Y линейная: $Y = a + b \cdot X$.

Чтобы количественно подтвердить или опровергнуть этот качественный вывод вычисляется выборочный коэффициент парной корреляции r_{xy} и проверяется его значимость по критерию Стьюдента.

Для повышения наглядности вычислений по МНК строится расчетная таблица 2.2.

Таблица 2.2.

№	x_i	y_i	x_i^2	$x_i y_i$	y_i^2	$\hat{y}_i$	e_i	e_i^2	A_i
1	4	11	16	44	121	11,408	-0,408	0,167	3,710
2	7	13	49	91	169	14,938	-1,938	3,754	14,904
3	5	12	25	60	144	12,585	-0,585	0,342	4,871
4	8	13	64	104	169	16,114	-3,114	9,697	23,954
5	6	15	36	90	225	13,761	1,239	1,535	8,260
6	7	15	49	105	225	14,938	0,063	0,004	0,417
7	9	17	81	153	289	17,290	-0,290	0,084	1,708
8	6	14	36	84	196	13,761	0,239	0,057	1,707
9	5	12	25	60	144	12,585	-0,585	0,342	4,871
10	7	15	49	105	225	14,938	0,063	0,004	0,417
11	9	17	81	153	289	17,290	-0,290	0,084	1,708
12	8	16	64	128	256	16,114	-0,114	0,013	0,712
13	7	15	49	105	225	14,938	0,063	0,004	0,417
14	8	17	64	136	289	16,114	0,886	0,785	5,212
15	7	19	49	133	361	14,938	4,063	16,504	21,382
16	9	18	81	162	324	17,290	0,710	0,503	3,942
Σ	112	239	818	1713	3651	239	0	33,879	98,192
среднее	7	14,938	51,125	107,063	228,188	14,938	0,000	2,117	6,137

С помощью данных таблицы 2.2 вычисляются значения числовых характеристик:

$$\text{cov}(x, y) = \overline{x \cdot y} - \bar{x} \cdot \bar{y} = 107,063 - 7 \cdot 14,938 = 2,500;$$

$$\text{var}(x) = \overline{x^2} - \bar{x}^2 = 51,125 - 7^2 = 2,125;$$

$$\text{var}(y) = \overline{y^2} - \bar{y}^2 = 228,188 - 14,938^2 = 5,059;$$

$$r_{xy} = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)} \cdot \sqrt{\text{var}(y)}} = \frac{2,500}{\sqrt{2,125} \cdot \sqrt{5,059}} = 0,763.$$

Значение выборочного коэффициента корреляции есть $r_{xy} = 0,763$. По формуле (2.2) вычисляется значение статистики t_r :

$$t_r = 0,763 \cdot \sqrt{\frac{14}{1 - 0,763^2}} = 0,763 \cdot \sqrt{\frac{14}{0,418}} = 4,410.$$

Выдвигается гипотеза $H_0: \rho_{xy} = 0$. Выбирается уровень значимости $\alpha = 0,05$, вычисляется $v = n - 2 = 16 - 2 = 14$ и по таблице распределения критерия Стьюдента определяется $t_{кр} = 2,145$.

Для $t_{кр}$ и t_r выполняется неравенство $t_r = 4,410 > t_{кр} = 2,145$ из которого делается вывод, что нулевая гипотеза должна быть отвергнута и необходимо признать, что между выборочными значениями переменных X и Y существует значимая линейная зависимость. Поэтому по формулам, приведенным ранее, вычисляются коэффициенты линейной зависимости a и b :

$$b = \frac{\text{cov}(x, y)}{\text{var}(x)} = \frac{2,500}{2,125} = 1,176;$$

$$a = \bar{y} - b \cdot \bar{x} = 14,938 - 1,176 \cdot 7 = 6,702;$$

Значит, построенное уравнение регрессии будет иметь вид:

$$\hat{y} = 6,702 + 1,176 \cdot x. \quad (2.23)$$

Далее определяется значимость коэффициентов $a = 6,702$ и $b = 1,176$ уравнения регрессии (2.23). Для чего вычисляются значения характеристик:

$$S^2 = \frac{n}{n-2} \cdot \text{Var}(e) = \frac{16}{14} \cdot 2,117 = 2,420;$$

$$S_a = \sqrt{\frac{S^2 \cdot x^2}{n \cdot \text{var}(x)}} = \sqrt{\frac{2,420 \cdot 51,125}{16 \cdot 2,125}} = 1,908;$$

$$S_b = \sqrt{\frac{S^2}{n \cdot \text{var}(x)}} = \sqrt{\frac{2,402}{16 \cdot 2,125}} = 0,267;$$

$$t_a = \frac{|a|}{S_a} = \frac{6,702}{1,908} = 3,514;$$

$$t_b = \frac{|b|}{S_b} = \frac{1,176}{0,267} = 4,410.$$

Затем задавшись уровнем значимости $\alpha = 0,05$; вычисляется число степеней свободы $\nu = n - 2 = 16 - 2 = 14$. И далее по таблице распределения критерия Стьюдента (приложение 1) определяется $t_{кр} = t(\alpha, \nu) = t(0,05; 14) = 2,145$.

Так как $t_a = 3,514 > t_{кр} = 2,145$, и $t_b = 4,410 > t_{кр} = 2,145$, то оба коэффициента значимо отличаются от нуля и должны быть оставлены в модели (2.23).

Далее определяется, насколько хорошо построенное уравнение регрессии (2.23) описывает наблюдаемые значения y . Для этого вычисляются числовые значения характеристик:

$$R^2 = r_{xy}^2 = 0,763^2 = 0,581;$$

$$F = \frac{R^2 \cdot (n-2)}{(1-R^2)} = \frac{0,581 \cdot 14}{0,419} = 19,447;$$

Задавшись уровнем значимости $\alpha = 0,05$; находятся по формулам: $k_1 = 1$, $k_2 = n - 2 = 16 - 2 = 14$ числа степеней свободы; далее по таблице распределения критерия Фишера – Снедекора (приложение 2) определяется $F_{кр} = F(\alpha, k_1, k_2) = F(0,05;1;14) = 4,600$. Так как $F = 19,447 > F_{кр} = 4,600$; то делается вывод, что построенное уравнение регрессии адекватно описывает наблюдаемые значения отклика y .

Затем вычисляется критерий средней относительной погрешности модели. Значение средней ошибки аппроксимации до 10% свидетельствует о хорошо подобранной модели уравнения.

$$A_i = \left| \frac{e_i}{y_i} \right| \cdot 100 ,$$

$$\bar{A} = \sum_{i=1}^{16} A_i / 16 = 98,192 / 16 = 6,137.$$

По критерию средней относительной погрешности построенная модель (2.23) так же адекватно описывает наблюдаемые значения отклика y , так как $6,137\% < 10,000\%$.

Для построенной модели значение коэффициента детерминации $R^2 = 0,581$; что свидетельствует о том, что 58,1% вариации значений результативной переменной y объясняется изменчивостью переменной x , а 41,9% вариации значений y объясняется воздействием случайного фактора.