ЭКОНОМЕТРИКА

Корреляционно-регрессионный анализ

Корреляционный анализ

- Метод, применяемый тогда, когда данные наблюдений или эксперимента можно считать случайными и выбранными из совокупности, распределенной по многомерному нормальному закону
- Основная задача корреляционного анализа - выявление связи между переменными путем точечной и интервальной оценки различных (парных, множественных, частных) коэффициентов корреляции

Коэффициент корреляции:

- **Формула:** $r = \dfrac{\overline{xy} - \overline{x}\,\overline{y}}{\sigma_x \sigma_y}$ $\quad |r| \le 1$
- **Оценка значимости с помощью *t*-критерия Стьюдента:**

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \ge t(\alpha, n-2)$$

- ***n-2* – число степеней свободы**
- ***α* – уровень значимости**

Пример для выборки из двухмерной нормально распределенной генеральной совокупности

- ***n = 122, r = 0,4, α = 0,05***
- **значение *t*-критерия Стьюдента:**

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0{,}4\sqrt{122-2}}{\sqrt{1-0{,}4^2}} = 4{,}78$$

- $t_{кр}\ (0{,}05;\ 120) = 1{,}98 < 4{,}78$
- ***r* – значимо отличается от 0**

Линейная парная регрессия

- **Рассмотрим зависимость между суточной выработкой продукции *Y* и величиной основных производственных фондов (ОПФ) *X* для совокупности 50 однотипных предприятий**

Вели-чина ОПФ	Середина интерв. $x_i \setminus y_i$	Суточная выработка продукции					Всего n_i	Гр. сред. y_i
		7-11	11-15	15-19	19-23	23-27		
		9	13	17	21	25		
20-25	22,5	2	1	-	-	-	3	10,3
25-30	27,5	3	6	4	-	-	13	13,3
30-35	32,5	-	3	11	7	-	21	17,8
35-40	37,5	-	1	2	6	2	11	20,3
40-45	42,5	-	-	-	1	1	2	23,0
Всего n_j		5	11	17	14	3	50	
Гр.сред. x_j		25,5	29,3	31,9	35,4	39,2		

Линейная парная регрессия: МНК

- **Метод наименьших квадратов:**

$$S = \sum_{i=1}^{l} (b_0 + b_1 x_i - \overline{y_i})^2 n_i \to \min.$$

- **Параметры уравнения регрессии находим из системы нормальных уравнений:**

$$\begin{cases} b_0 n + b_i \sum_{i=1}^{l} x_i n_i = \sum_{i=1}^{l} \overline{y_i} n_i, \\ b_0 \sum_{i=1}^{l} x_i n_i + b_1 \sum_{i=1}^{l} x_i^2 n_i = \sum_{i=1}^{l} x_i \overline{y_i} n_i. \end{cases}$$

- **Уравнение регрессии:**

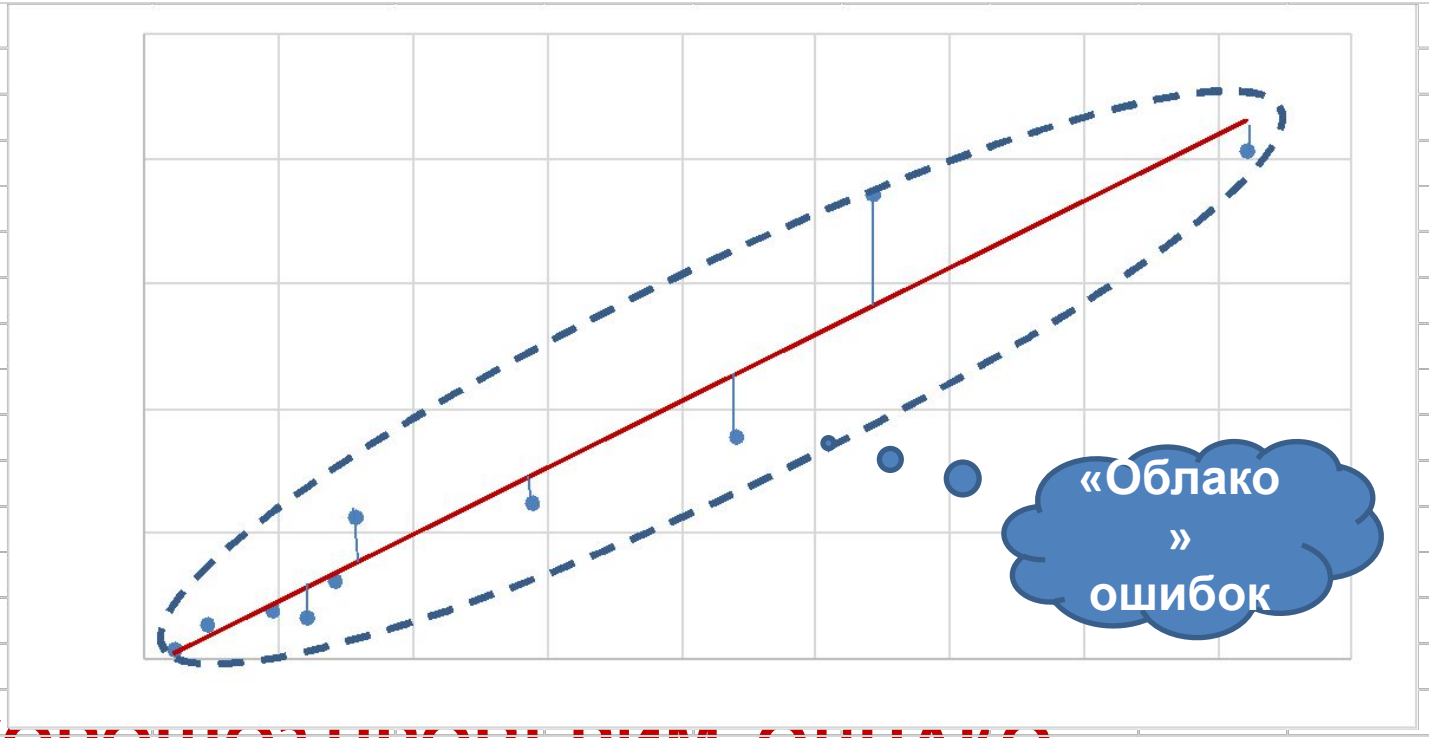
$$y_x = 0,6762x - 4,79$$

- **Коэффициент регрессии: $b_1 = 0,6762$ означает, что при увеличении ОПФ на 1 ед. суточная выработка предприятия увеличивается в среднем на 0,6762 ед.**

Пример «плохой» модели парной линейной регрессии (Excel)

- По данным, полученным от 10 предприятий, изучается зависимость объема выпуска продукции *Y* (млн. руб.) от численности производственного персонала *X* (чел.)
- Графически зависимость вроде бы близка к линейной

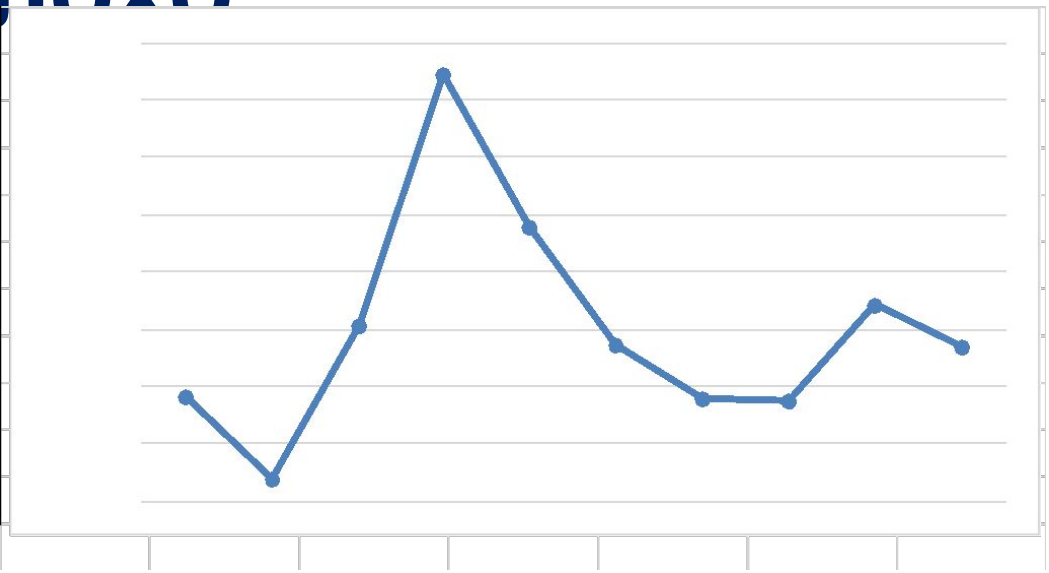
№	X	Y
1	2 890	62 240
2	4 409	88 569
3	210	3 118
4	5 436	186 256
5	1 559	56 262
6	940	19216
7	1 197	16 567
8	8 212	203 456
9	459	13 425
10	1 405	31 163
ЛИНЕЙН		
26,70618	**-3323,694**	
2,667583	9682,778	
0,926082	20727,36	



ВСЁ ХОРОШО? ПРОВЕРИМ, ОДНАКО…

Анализ остатков: всё очень плохо

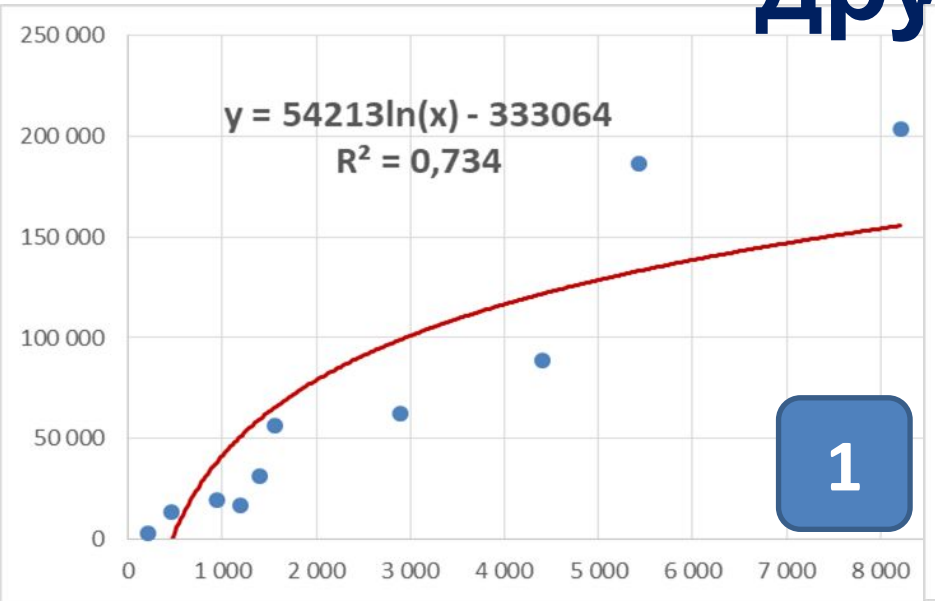
№	X	Y	Yмод	e
1	2 890	62 240	73 856,64	-11 616,64
2	4 409	88 569	114 423,05	-25 854,05
3	210	3 118	2 284,56	833,44
4	5 436	186 256	141 850,12	44 405,88
5	1 559	56 262	38 310,95	17 951,05
6	940	19216	21 779,94	-2 563,94
7	1 197	16 567	28 643,38	-12 076,38
8	8 212	203 456	215 985,97	-12 529,97
9	459	13 425	8 934,35	4 490,65
10	1 405	31 163	34 198,23	-3 035,23

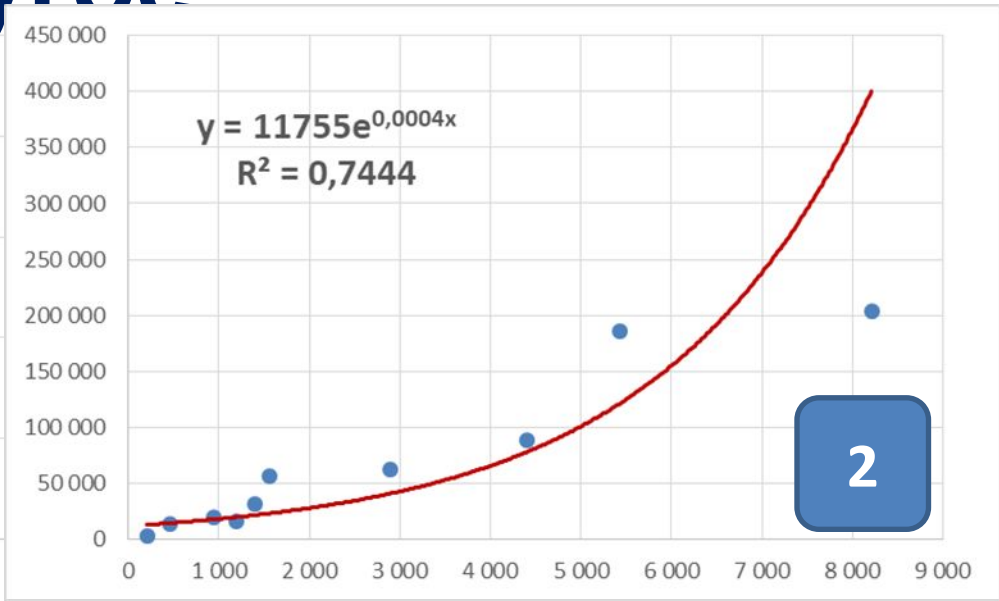


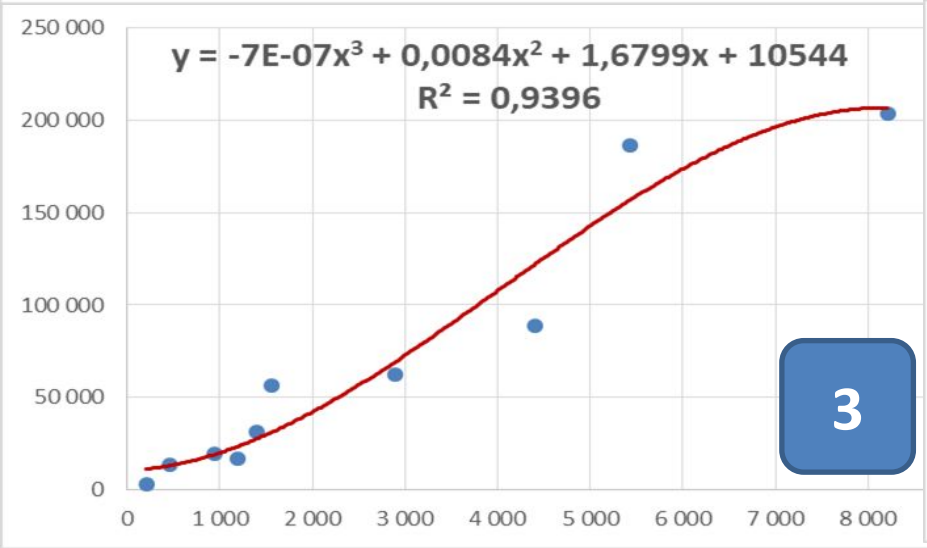
- Среднее значение остатков: $e_{\text{ср}} = \frac{\sum_{i=1}^{n} e_i}{n} = 0,48$
- Отклонение от линии регрессии: $S_e = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}} = 20727,36$
- Средняя относительная ошибка аппроксимации:

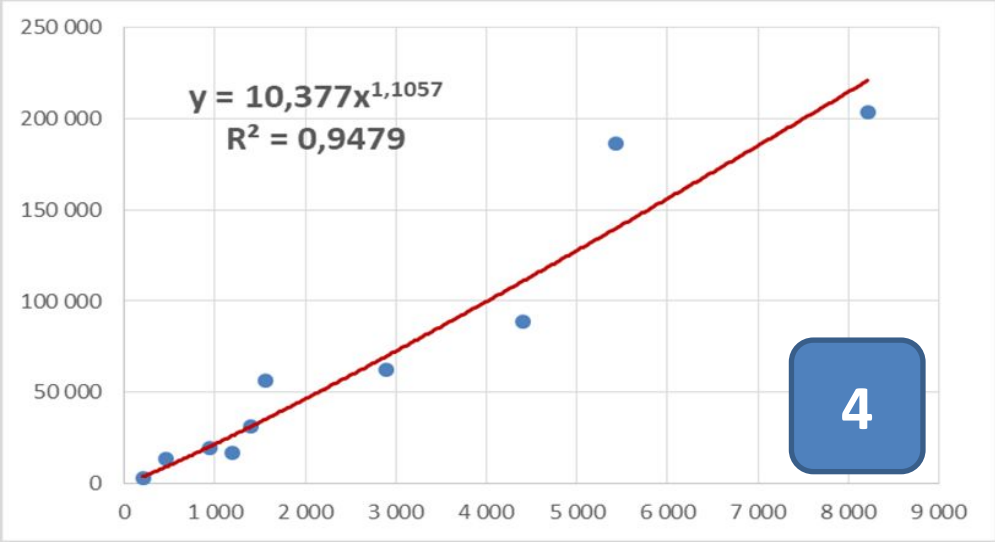
$$\bar{e}_{\text{отн}} = \frac{1}{n}\sum_{i=1}^{n} \frac{|e_i|}{y_i} \times 100\% = 26,6\%$$ - ОЧЕНЬ БОЛЬШАЯ ОШИБКА!

Попробуем подобрать что-то другое?









1 – логарифмическая, 2 – экспоненциальная, 3 – кубическая, 4 - степенная

Регрессионный анализ

Предназначен для исследования зависимости исследуемой переменной (признака) Y от различных факторов $X_1, X_2, \dots, X_m$ и отображения их взаимосвязи в форме регрессионной модели:

$$\widehat{Y} = f(X_1, X_2, \dots, X_m),$$

которая показывает, каково будет в среднем значение Y, если X_j ($j = 1, \dots, m$) примут конкретные значения

При этом $Y_i = \widehat{Y}_i + \varepsilon_i$ $(i = 1, \dots, n)$, где n – объем выборки, ε_i - остатки (остаточная компонента)

Для линейной парной регрессии имеем:

$$y_i = a + bx_i + \varepsilon_i$$

Условия Гаусса-Маркова и свойства оценок МНК

- Математическое ожидание: $M(\varepsilon_i) = 0$
- Случайный характер остатков ε_i
- Независимость остатков ε_i и ε_j $(i \neq j)$ (отсутствие автокорреляции в остатках)
- Постоянство дисперсии: $D(\varepsilon_i) = \sigma_\varepsilon^2 = Const$
- Нормальный характер распределения остатков ε_i
- При выполнении этих условий оценки коэффициентов регрессии, полученные МНК, будут обладать следующими свойствами:
  - несмещенности (математическое ожидание остатков = 0)
  - эффективности (характеризуются наименьшей дисперсией)
  - состоятельности (улучшение точности с ростом *n*)

Пример построения линейной модели множественной регрессии

- Требуется провести корреляционно-регрессионный анализ статистических данных объема реализации продукции фирмы *Y* за 16 месяцев для целей дальнейшего прогнозирования
- Факторы: X_1 - время, X_2 - затраты на рекламу, X_3 - цена товара, X_4 - средняя цена у конкурентов, X_5 - индекс потребительских цен
- Регрессионную модель первоначально будем рассматривать в виде:

$$\widehat{Y} = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + b_3 \cdot X_3 + b_4 \cdot X_4 + b_5 \cdot X_5$$

Исходные данные:

Y	X1	X2	X3	X4	X5	Y	X1	X2	X3	X4	X5
126	1	4,0	15,0	17,0	100,0	367	9	19,8	15,8	18,2	108,3
137	2	4,8	14,8	17,3	98,4	367	10	10,6	16,9	16,8	109,2
148	3	3,8	15,2	16,8	101,2	321	11	8,6	16,3	17,0	110,1
191	4	8,7	15,5	16,2	103,5	307	12	6,5	16,1	18,3	110,7
274	5	8,2	15,5	16,0	104,1	331	13	12,6	15,4	16,4	110,3
370	6	9,7	16,0	18,0	107,0	345	14	6,5	15,7	16,2	111,8
432	7	14,7	18,1	20,2	107,4	364	15	5,8	16,0	17,7	112,3
445	8	18,7	13,0	15,8	108,5	384	16	5,7	15,1	16,2	112,9

Что говорят временны*е* графики?

Корреляционная матрица *R*

	Y	X1	X2	X3	X4	X5
Y	1					
X1	0,677985	1				
X2	0,645918	0,106455	1			
X3	0,232895	0,173717	-0,00335	1		
X4	0,226319	-0,051	0,204043	0,697751	1	
X5	0,816018	0,960204	0,273373	0,235428	0,030784	1

- Факторы X_3 и X_4 - слабо связаны с признаком Y ($r_{3Y} = 0,233;\ r_{4Y} = 0,226$)
- Сильная связь между X_1 и X_5 ($r_{15} = 0,96$)
- Т.к. $r_{1Y} = 0,678 < r_{5Y} = 0,816$, то исключаем фактор X_1, а X_5 - оставляем

Мультиколлинеарность

	X2	X3	X4	X5
X2	1	-0,00335	0,204043	0,273373
X3	-0,00335	1	0,697751	0,235428
X4	0,204043	0,697751	1	0,030784
X5	0,273373	0,235428	0,030784	1

Определитель матрицы R_1 вычислим с помощью функции МОПРЕД(*массив*)

$$det(R_1) = 0,373$$

- Значение статистики Фаррара-Глоубера:

$$FG_{\text{набл}} = -\left(n - 1 - \frac{1}{6} \cdot (2k + 5)\right) \cdot ln\left(det(R_1)\right)$$

$$= -\left(15 - \frac{13}{6}\right) \cdot ln(0,373) = 12,66$$

- На уровне значимости $\alpha = 0,05$ и количестве степеней свободы $\frac{k(k-1)}{2} = 6$ при помощи функции ХИ2.ОБР.ПХ найдем табличное значение $FG_{\text{крит}} = \chi^2 = 12,59$
- Т.к. $FG_{\text{набл}} > FG_{\text{крит}}$, то в массиве объясняющих переменных есть мультиколлинеарность

Выявление коллинеарных факторов

	X2	X3	X4	X5
X2	1,251803	0,54356	-0,6208	-0,45107
X3	0,54356	2,375858	-1,74853	-0,65411
X4	-0,6208	-1,74853	2,331021	0,509606
X5	-0,45107	-0,65411	0,509606	1,261618

С помощью функции МОБР(*массив*) найдем обратную матрицу $C = R_1^{-1}$

- Вычислим значения *F*-критерия:

$$F_j = (c_{jj} - 1) \cdot \frac{n - k - 1}{k}$$

F2	F3	F4	F5
0,692	3,784	3,660	0,719

- Табличное значение $F_{\text{табл}} = 3,357$
- Факторы X_3 и X_4 - коллинеарны с другими факторами

- Находим частные коэффициенты корреляции:

$$r_{ij(\dots)} \frac{-c_{ij}}{\sqrt{c_{ii} \cdot c_{jj}}}$$

$$r_{2,3(4,5)} = -0,315;\ r_{2,4(3,5)} = 0,363;\ r_{2,5(3,4)} = 0,359;$$
$$r_{3,4(2,5)} = 0,743;\ r_{3,5(2,4)} = 0,378;\ r_{4,5(2,3)} = -0,297$$

- Рассчитываем значения *t*-критерия Стьюдента:

$$t_{ij} = \frac{r_{ij(\dots)} \cdot \sqrt{n-k-1}}{\sqrt{1-r_{ij(\dots)}^2}}$$

$$t_{2,3} = -1,102;\ t_{2,4} = 1,293;\ t_{2,5} = 1,275;\ t_{3,4} = 3,682;$$
$$t_{3,5} = 1,353;\ t_{4,5} = -1,032$$

- $t_{\text{табл}}(0,05; 11) = 2,201$ (функция СТЬЮДЕНТ.ОБР.2Х)
- Для факторов X_3 и X_4: $|t_{3,4}| > |t_{\text{табл}}|$ и $r_{3,4(2,5)} = 0,743 \sim 1$
- Факторы X_3 и X_4 - коллинеарны. Какой из них исключить?
- Удаляем из модели X_3, т.к. $F_3 = 3,784 > F_4 = 3,660$

Вернемся к временны*м* графикам:

Сходство графиков

X_3 и X_4 - графики практически идентичны

X_1 и X_5 - линейная зависимость

«Длинная» и «короткая» регрессии

- «Длинная» регрессия с факторами X_2, X_4, X_5:

Дисперсионный анализ					
	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	3	138 429,778	46 143,259	27,292	1,20724E-05
Остаток	12	20 288,659	1 690,722		
Итого	15	158 718,438			

	Коэффициенты	*Станд.ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	-1654,763	306,264	-5,403	0,000	-2322,054	-987,472
X2	9,052	2,295	3,945	0,002	4,052	14,051
X4	10,539	9,521	1,107	0,290	-10,206	31,284
X5	15,825	2,447	6,468	0,000	10,494	21,156

$$\hat{Y} = -1654,76 + 9,05 \cdot X_2 + 10,54 \cdot X_4 + 15,83 \cdot X_5;\ (R^2 = 0,872)$$

- «Короткая» регрессия с факторами X_2, X_5:

Дисперсионный анализ					
	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	2	136 358,334	68 179,167	39,639	2,93428E-06
Остаток	13	22 360,104	1 720,008		
Итого	15	158 718,438			

	Коэффициенты	*Станд.ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	-1 471,314	259,766	-5,664	0,000	-2 032,505	-910,124
X2	9,568	2,266	4,223	0,001	4,673	14,464
X5	15,753	2,467	6,386	0,000	10,424	21,082

$$\hat{Y} = -1471,31 + 9,57 \cdot X_2 + 15,75 \cdot X_5;\ (R^2 = 0,859)$$

Какую регрессию выбрать?

- **Вычислим *F*-статистику:**

$$F_{\text{набл}} = \frac{(ESS_{\text{кор}} - ESS_{\text{длин}})/q}{ESS_{\text{длин}}/(n-k-1)}$$

$$= \frac{(22360,104 - 20288,659)/1}{20288,659/(16-3-1)} = 1,225$$

- **Сравним с** $F_{\text{табл}}(0,05; 1; 12) = 4,747$
- **Если** $F_{\text{набл}} > F_{\text{табл}}$**, то выбираем «длинную» регрессию, в противном случае – «короткую»**
- **В нашем случае выбираем «короткую»:**

$$\widehat{Y} = -1471,31 + 9,57 \cdot X_2 + 15,75 \cdot X_5$$