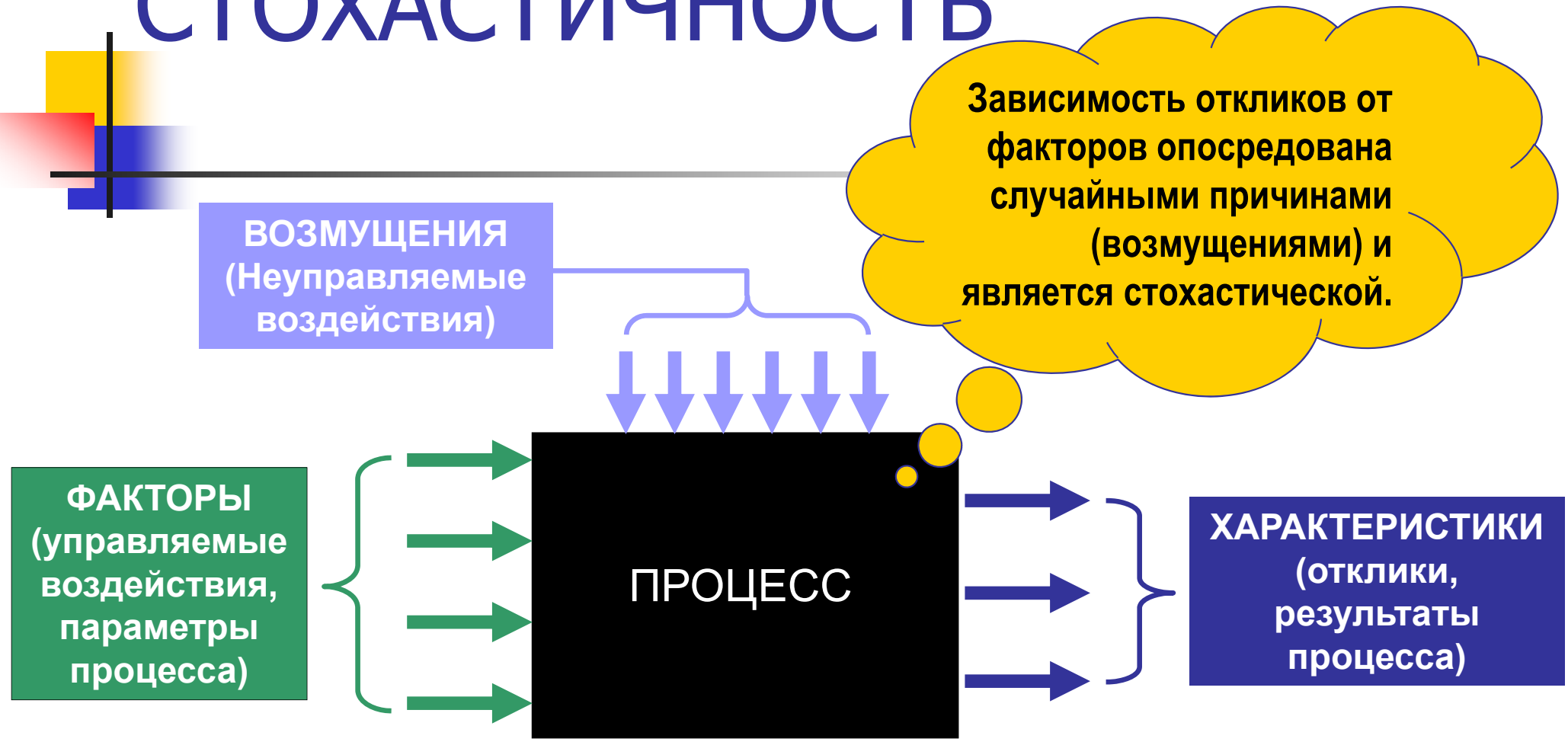
ОБРАБОТКА И АНАЛИЗ ЧИСЛОВОЙ ИНФОРМАЦИИ

ПАРНЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ

© Румянцев Михаил Игоревич,
профессор, канд. техн. наук

Парный регрессионный анализ –
это метод математической статистики,
который позволяет найти
наиболее точное и достоверное отображение
(модель, аппроксимацию)
стохастической зависимости
между откликом и одним из факторов

СТОХАСТИЧНОСТЬ



Стохастичность зависимости проявляется, например, в том, что при одних и тех же значениях факторов в различные моменты времени будут обнаружены различные значения отклика

МОДЕЛЬ ПАРНОГО РЕГРЕССИОННОГО АНАЛИЗА

При стохастической связи между параметрами некоторого значения отклика с соответствующим значением фактора может быть представлена в виде двух составляющих:

$$y_i = \varphi(x_i) + \varepsilon_i$$

$\varphi(x_i)$ - систематическая (объясненная) составляющая. Она обусловлена существованием зависимости между откликом и фактором.

ε_i - случайная составляющая. Она обусловлена разнообразными возмущениями и вызывает отклонения y_i от соответствующих реальной зависимости.

ПОСЛЕДОВАТЕЛЬНОСТЬ РЕГРЕССИОННОГО АНАЛИЗА

1. Определить вид уравнения регрессии
2. Оценить допустимость отображения исследуемой зависимости выбранным уравнением регрессии

ОПРЕДЕЛЕНИЕ ВИДА УРАВНЕНИЯ

Необходимо определить вид систематической составляющей.

Т.к. используются выборки ограниченного объема, действительная связь между откликом и фактором представляется оценкой (отображением) этой связи

$$\boxed{\hat{y} = \hat{\varphi}(x)} \approx y = \varphi(x)$$

Уравнение регрессии

ЧТО ПРЕДСТАВЛЯЕТ СОБОЙ УРАВНЕНИЕ РЕГРЕССИИ?

Уравнение регрессии – статистическое отображение объясненной составляющей взаимосвязи между параметрами

ЛИНИЯ РЕГРЕССИИ

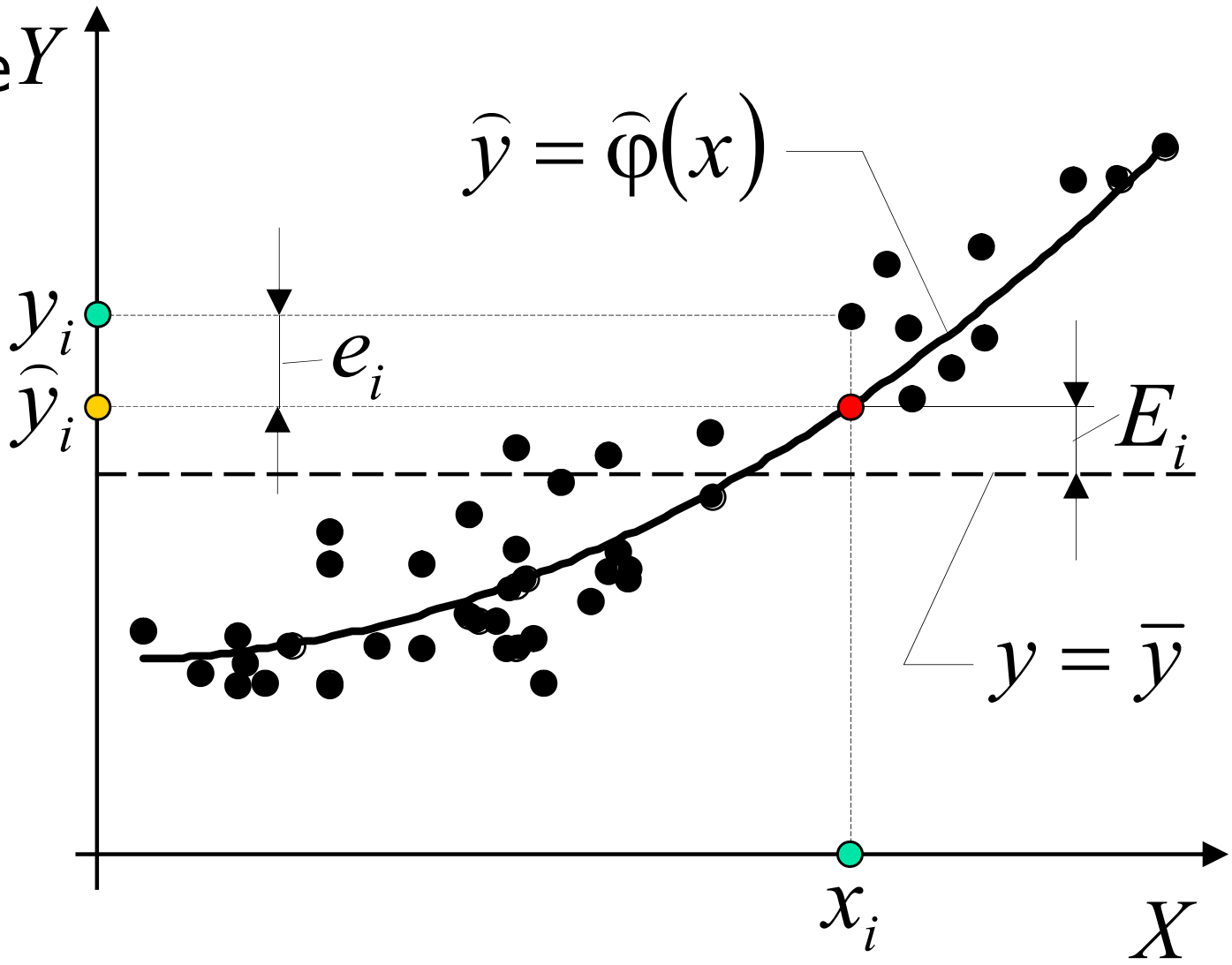
y_i - фактическое значение отклика

$\widehat{y}_i$ - оценка значения отклика

$\overline{y}_i$ - среднее выборочное значение отклика

e_i - случайное отклонение

E_i - объясненное отклонение



ПРИНЦИП НАИМЕНЬШИХ КВАДРАТОВ

Наилучшей оценкой исследуемой зависимости является та, которая дает наименьшую сумму квадратов отклонений наблюдаемых значений отклика от рассчитанных по уравнению регрессии при тех же значениях фактора

$$\sum_{i=1}^{n} e_i^{\,2} = \sum_{i=1}^{n} \left(y_i - \widehat{y}_i\right)^2 = min$$

ЛИНЕЙНАЯ АППРОКСИМАЦИЯ

ИСТИННАЯ ЗАВИСИМОСТЬ

$$y = \beta_0 + \beta_1 x$$

ИСТИННЫЕ КОЭФФИЦИЕНТЫ

УРАВНЕНИЕ РЕГРЕССИИ

$$\hat{y} = b_0 + b_1 x$$

КОЭФФИЦИЕНТЫ РЕГРЕССИИ

$$b_0 \approx \beta_0$$

$$b_1 \approx \beta_1$$

ОТРЕЗОК(<Y>;<X>)

НАКЛОН(<Y>;<X>)

СОСТАВЛЯЮЩИЕ КАЧЕСТВА АППРОКСИМАЦИИ

ОСТАТОЧНАЯ ДИСПЕРСИЯ

$$S_e^2 = \frac{1}{n-k}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

Характеризует рассеяние фактических значений отклика относительно линии регрессии вследствие случайных причин

ОБЪЯСНЕННАЯ ДИСПЕРСИЯ

$$S_E^2 = \frac{1}{k-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$$

Характеризует рассеяние значений отклика относительно его среднего выборочного значения вследствие существования исследуемой взаимосвязи

ОЦЕНКА КАЧЕСТВА АППРОКСИМАЦИИ СРАВНЕНИЕМ ДИСПЕРСИЙ

Связь между параметрами может быть отображена данным уравнением регрессии, если объясненная дисперсия существенно больше остаточной

ТЕСТ ФИШЕРА

$$F_p = \frac{S_E^2}{S_e^2} > \underbrace{F[\alpha; k-1; n-k]}_{\text{FРАСПОБР}(\alpha;k-1;n-k)}$$

ОЦЕНКА КАЧЕСТВА АППРОКСИМАЦИИ С ПОМОЩЬЮ ПОКАЗАТЕЛЯ R^2

Связь между параметрами может быть отображена данным уравнением регрессии, если выполняется тест Фишера

ПОКАЗАТЕЛЬ ДОСТОВЕРНОСТИ АППРОКСИМАЦИИ

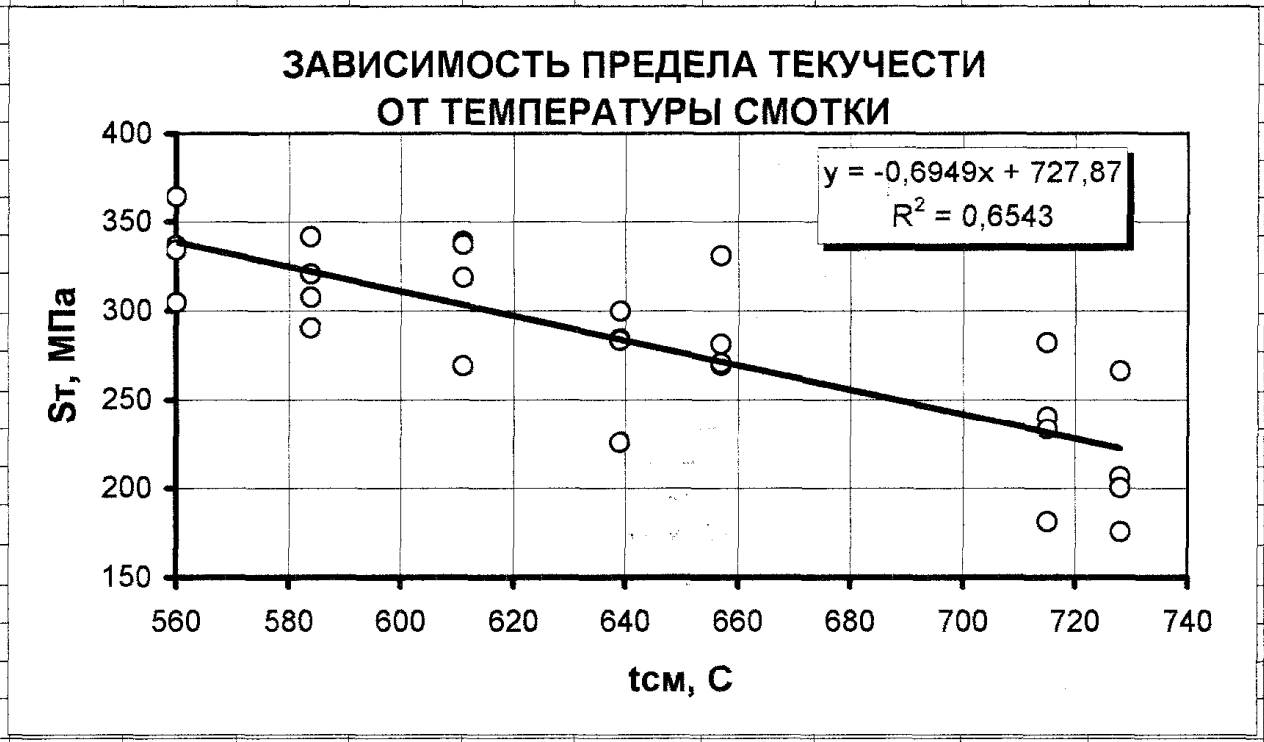
$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \frac{n-1}{n-k}$$

ТЕСТ ФИШЕРА

$$F_p = \frac{R^2}{1-R^2} \frac{n-k}{k-1} > F[\alpha; k-1; n-k]$$

ПОСТРОЕНИЕ АППРОКСИМАЦИЙ В MS EXCEL С ПРИМЕНЕНИЕМ ИНСТРУМЕНТА «ЛИНИЯ ТРЕНДА»

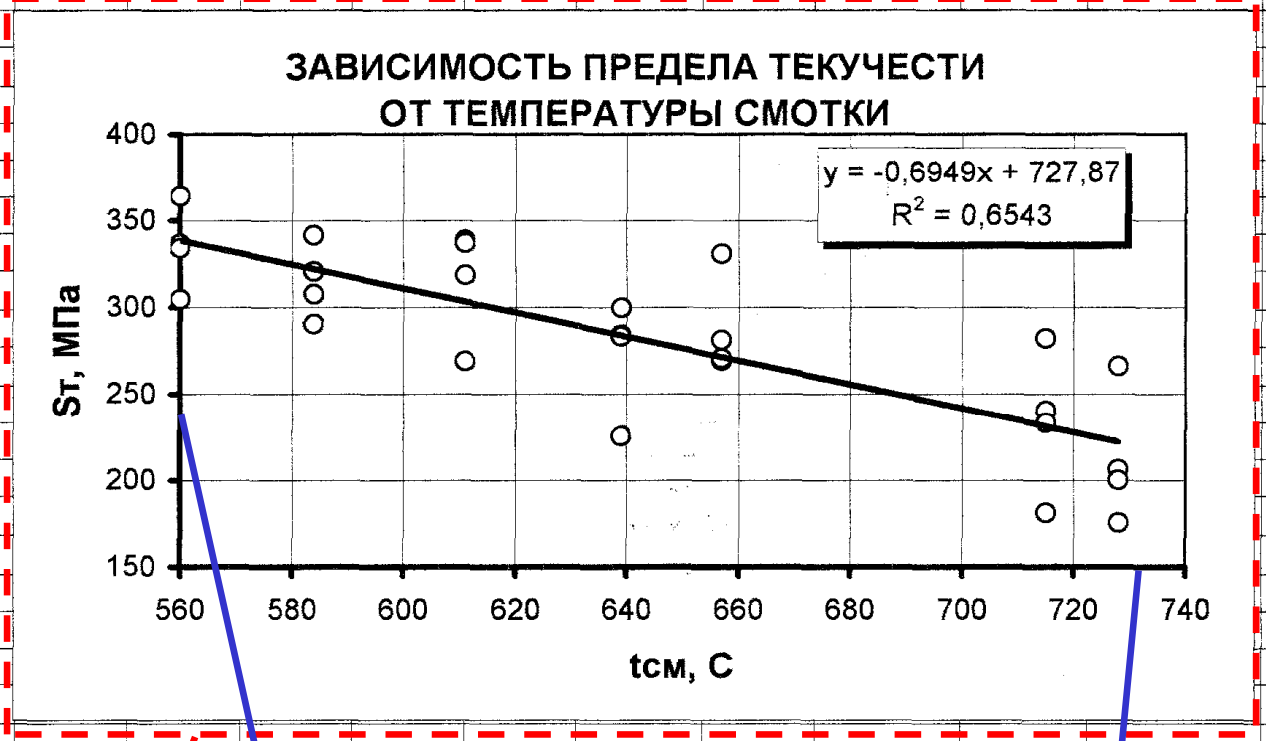
	A	B	C	D
1	ИСХОДНЫЕ ДАННЫЕ			
2	i	Sт, МПа	tкп,С	tсм,С
3	1	364,0	788	560
4	2	341,2	788	584
5	3	338,4	788	611
6	4	284,1	788	639
7	5	330,9	788	657
8	6	281,9	788	715
9	7	265,9	788	728
10	8	336,4	834	560
11	9	320,5	834	584
12	10	336,8	834	611
13	11	299,4	834	639
14	12	269,3	834	657
15	13	240,3	834	715
16	14	206,3	834	728
17	15	334,2	875	560
18	16	290,3	875	584
19	17	318,5	875	611
20	18	283,2	875	639
21	19	281,0	875	657
22	20	233,1	875	715
23	21	200,0	875	728
24	22	304,4	917	560
25	23	307,5	917	584
26	24	268,8	917	611
27	25	225,4	917	639
28	26	270,3	917	657
29	27	181,3	917	715
30	28	175,3	917	728
31				



	n=	28	R2	k	Fp	Fт	ВЫВОД
y = -0,6949x + 727,87			0,6540	2	29,7125	4,2100	Надежная
y = -445,12Ln(x) + 3157,4			0,6474	2	28,9778	4,2100	Надежная
y = 1E+07x^(-1,6836)			0,6256	2	26,7252	4,2100	Надежная
y = 1501,8e^(-0,0026x)			0,6348	2	27,6449	4,2100	Надежная
y = -0,0017x2 + 1,4773x + 30,696			0,6631	3	14,7935	3,3541	Надежная

Рассматривается взаимосвязь отклика с фактором, для которого коэффициент парной корреляции наибольший

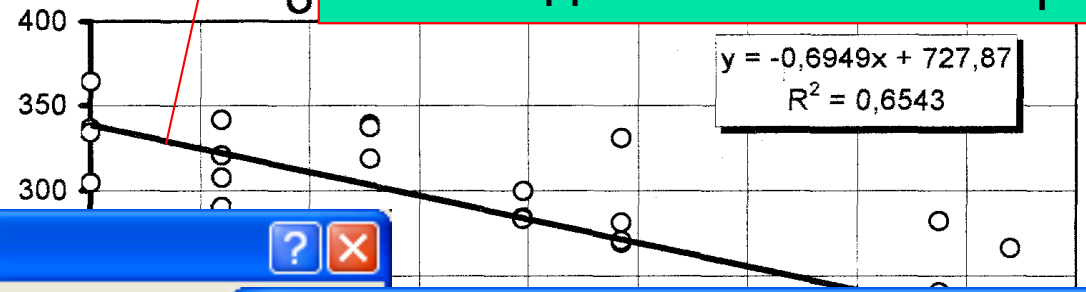
	A	B	C	D
1	ИСХОДНЫЕ ДАННЫЕ			
2	i	Sт, МПа	tкп,С	tсм,С
3	1	364,0	788	560
4	2	341,2	788	584
5	3	338,4	788	611
6	4	284,1	788	639
7	5	330,9	788	657
8	6	281,9	788	715
9	7	265,9	788	728
10	8	336,4	834	560
11	9	320,5	834	584
12	10	336,8	834	611
13	11	299,4	834	639
14	12	269,3	834	657
15	13	240,3	834	715
16	14	206,3	834	728
17	15	334,2	875	560
18	16	290,3	875	584
19	17	318,5	875	611
20	18	283,2	875	639
21	19	281,0	875	657
22	20	233,1	875	715
23	21	200,0	875	728
24	22	304,4	917	560
25	23	307,5	917	584

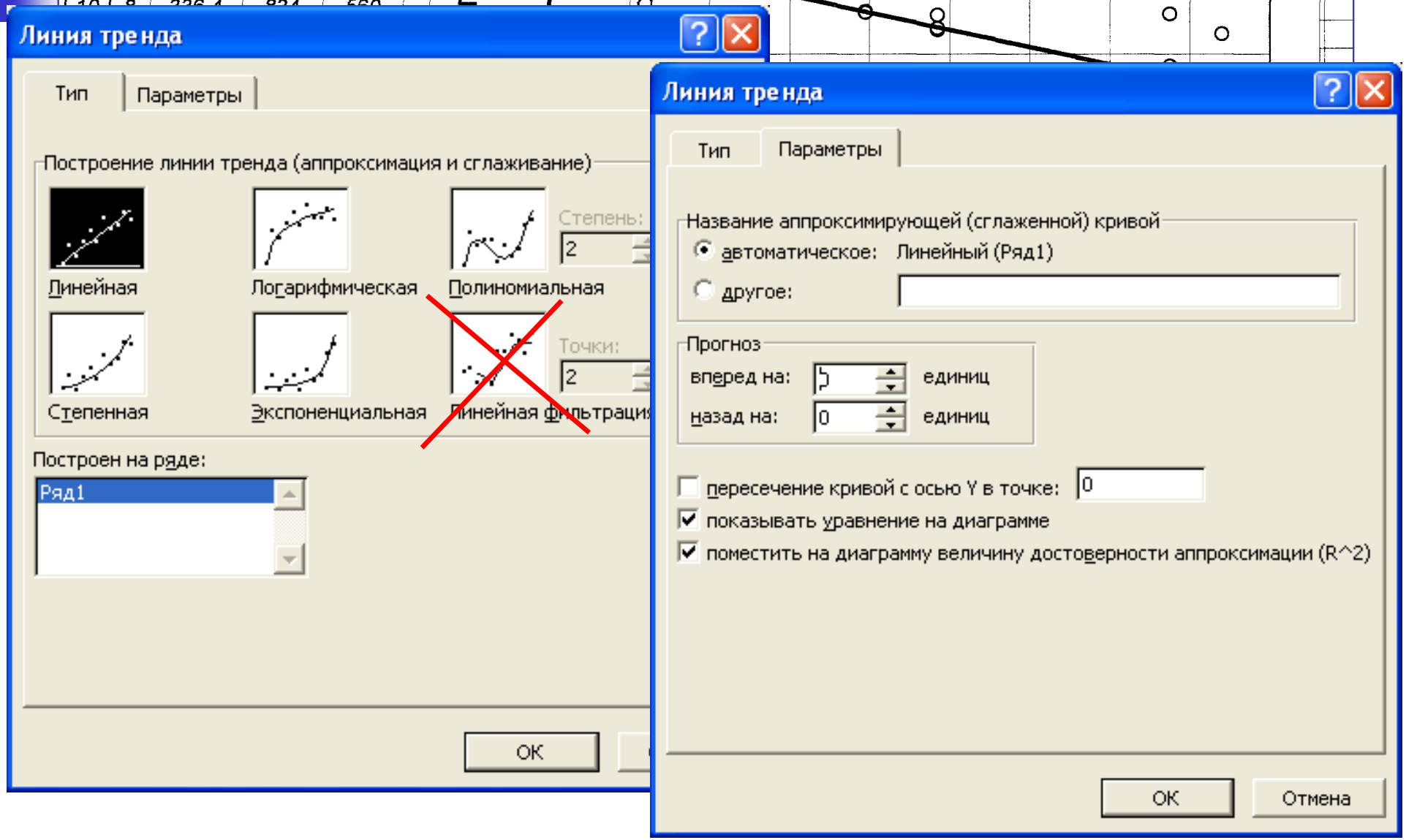


| | | |

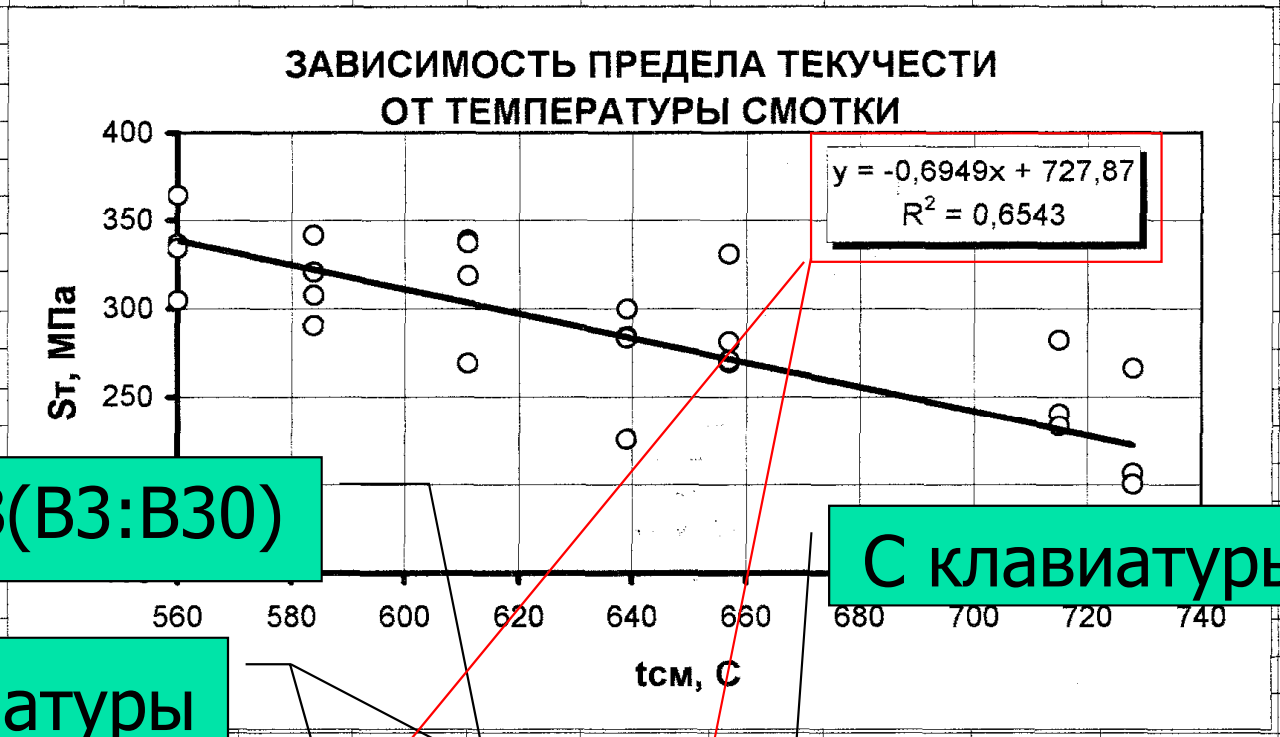
После построения и редактирования диаграммы добавить линию тренда

	A	B	C	D
1	ИСХОДНЫЕ ДАННЫЕ			
2	i	Sт, МПа	tкп,С	tсм,С
3	1	364,0	788	560
4	2	341,2	788	584
5	3	338,4	788	611
6	4	284,1	788	639
7	5	330,9	788	657
8	6	281,9	788	715
9	7	265,9	788	728





	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	ИСХОДНЫЕ ДАННЫЕ														
2	i	Sт, МПа	tкп,С	tсм,С											
3	1	364,0	788	560											
4	2	341,2	788	584											
5	3	338,4	788	611											
6	4	284,1	788	639											
7	5	330,9	788	657											
8	6	281,9	788	715											
9	7	265,9	788	728											
10	8	336,4	834	560											
11	9	320,5	834	584											
12	10	336,8	834	611											
13	11	299,4	834	639											
14	12	269,3													
15	13	240,3													
16	14	206,3													
17	15	334,2	875	560											
18	16	290,3	875	584											
19	17	318,5													
20	18	283,2													
21	19	281,0													
22	20	233,1	875	715											
23	21	200,0	875	728		y = -0,6949x + 727,87									
24	22	304,4	917	560		y = -445,12Ln(x) + 3157,4									
25	23	307,5	917	584		y = 1E+07x^(-1,6836)									
29	27	181,3	917	715											



	n=	28	R2	k	Fр	Fт	ВЫВОД
			0,6540	2	29,7125	4,2100	Надежная
			0,6474	2	28,9778	4,2100	Надежная
			0,6256	2	26,7252	4,2100	Надежная
			0,6348	2	27,6449	4,2100	Надежная
			0,6631	3	14,7935	3,3541	Надежная

=СЧЁТЗ(B3:B30)

С клавиатуры

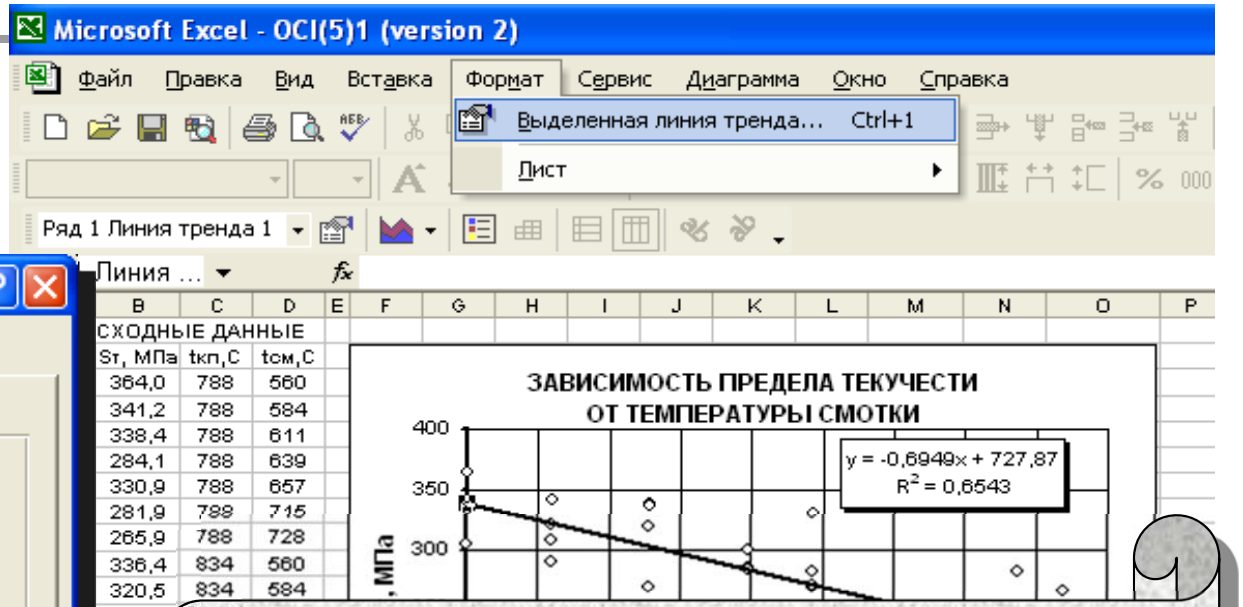
С клавиатуры

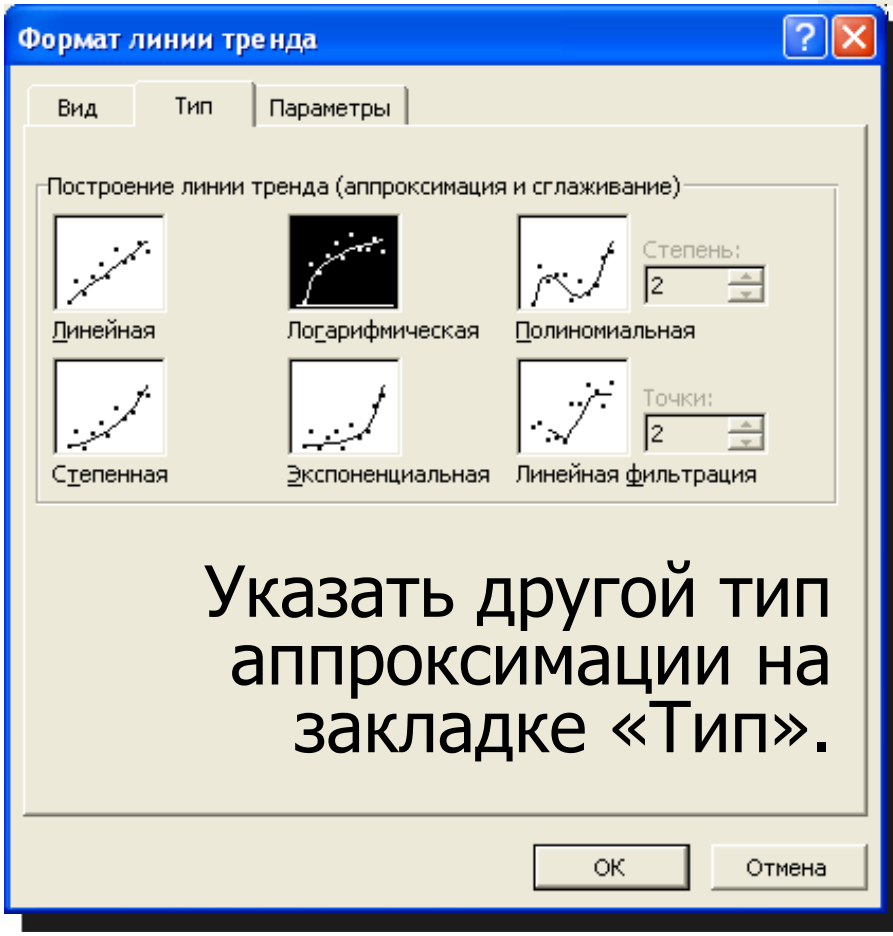
=K23*(J22-L23)/((1-K23)*(L23-1))

=FРАСПОБР(0,05;L23-1;J22-1)

=ЕСЛИ(M23>N23;"Надежная";"Не надежная")

Для получения новой аппроксимации форматировать линию тренда





Указать другой тип аппроксимации на закладке «Тип».

Если каждый из вариантов аппроксимации наносить на диаграмму как добавление линии тренда, то на диаграмме будут одновременно отображены все рассмотренные варианты.

Связь между какими величинами анализировалась?

Анализировалась связь между пределом текучести металла σ_T и температурой смотки tсм при прокатке на ШСГП.

Сколько и какие аппроксимации были рассмотрены ?

Рассмотрели пять аппроксимаций:

$\sigma_T = 0{,}6949\, t_{см} + 727{,}87$	(1)	$F_p = 49{,}1445$	$F_{95} = 4{,}2100$
$\sigma_T = 3157{,}4 - 445{,}12\, ln\left(t_{см}\right)$	(2)	$F_p = 47{,}7379$	$F_{95} = 4{,}2100$
$\sigma_T = 10^7\, t_{см}^{-1{,}6836}$	(3)	$F_p = 43{,}4444$	$F_{95} = 4{,}2100$
$\sigma_T = 1501{,}8\, e^{-0{,}0026 t_{см}}$	(4)	$F_p = 45{,}1939$	$F_{95} = 4{,}2100$
$\sigma_T = 30{,}696 + 1{,}4773\, t_{см} - 0{,}0017\, t_{см}^2$	(5)	$F_p = 24{,}6030$	$F_{95} = 3{,}3541$

Какие из рассмотренных аппроксимаций являются статистически значимыми?

С доверительной вероятностью 95% статистически значимыми являются все рассмотренные аппроксимации, т. к. во всех случаях рассчитанные числа Фишера больше табличных.

Какая из аппроксимаций является наилучшим отображением связи между параметрами?

Наилучшим отображением связи между пределом текучести металла и температурой смотки является линейная аппроксимация

$$\sigma_T = 727{,}87 - 0{,}6949\, t_{см}$$

$$\left(R^2 = 0{,}6540;\ F_p = 49{,}1445;\ F_{95} = 4{,}2100\right)$$