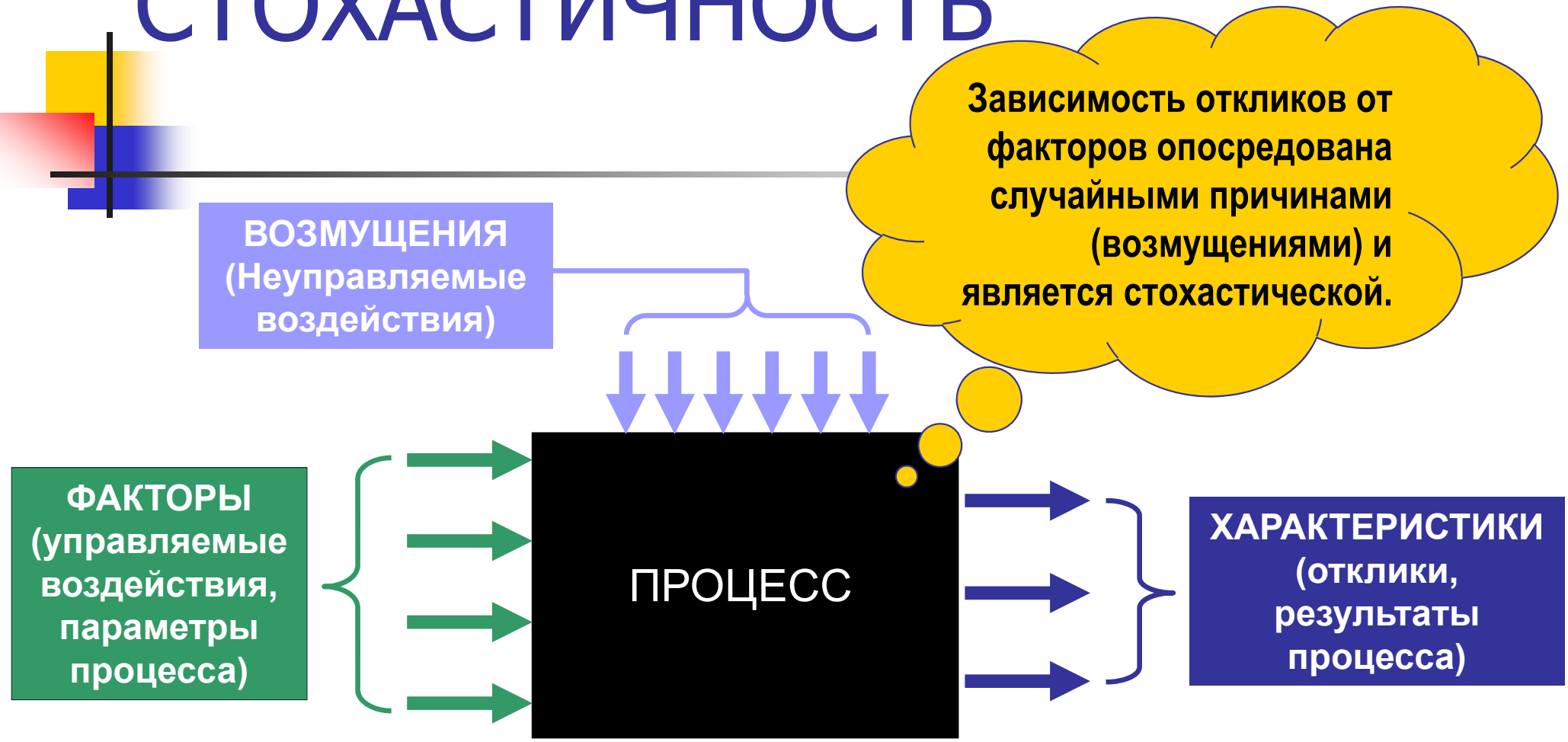
ОБРАБОТКА И АНАЛИЗ ЧИСЛОВОЙ ИНФОРМАЦИИ

МНОЖЕСТВЕННЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ

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

Множественный регрессионный анализ – это метод математической статистики, который позволяет найти наиболее точное и достоверное отображение (модель, аппроксимацию) стохастической зависимости между откликом и несколькими одновременно действующими факторами

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 МНОЖЕСТВЕННОГО РЕГРЕССИОННОГО АНАЛИЗА

При стохастической связи между параметрами зависимость некоторого значения отклика от соответствующих значений факторов может быть представлена в виде двух составляющих:

$$y_i = \varphi\left(x_{1i}, \ldots, x_{ji}, \ldots, x_{mi}\right) + \varepsilon_i$$

Систематическая (объясненная) составляющая. Обусловлена существованием зависимости между откликом и факторами.

Случайная составляющая. Обусловлена разнообразными возмущениями и вызывает отклонения отклика от значений, 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 факторами представляется оценкой (отображением) этой связи

$$\boxed{\hat{y}_i = \hat{\varphi}\left(x_{1i},\ldots,x_{ji},\ldots,x_{mi}\right)} \approx \varphi\left(x_{1i},\ldots,x_{ji},\ldots,x_{mi}\right)$$

Уравнение регрессии

ЧТО ПРЕДСТАВЛЯЕТ СОБОЙ УРАВНЕНИЕ РЕГРЕССИИ?

Уравнение регрессии – статистическое отображение объясненной составляющей взаимосвязи между параметрами

ЛИНИЯ РЕГРЕССИИ

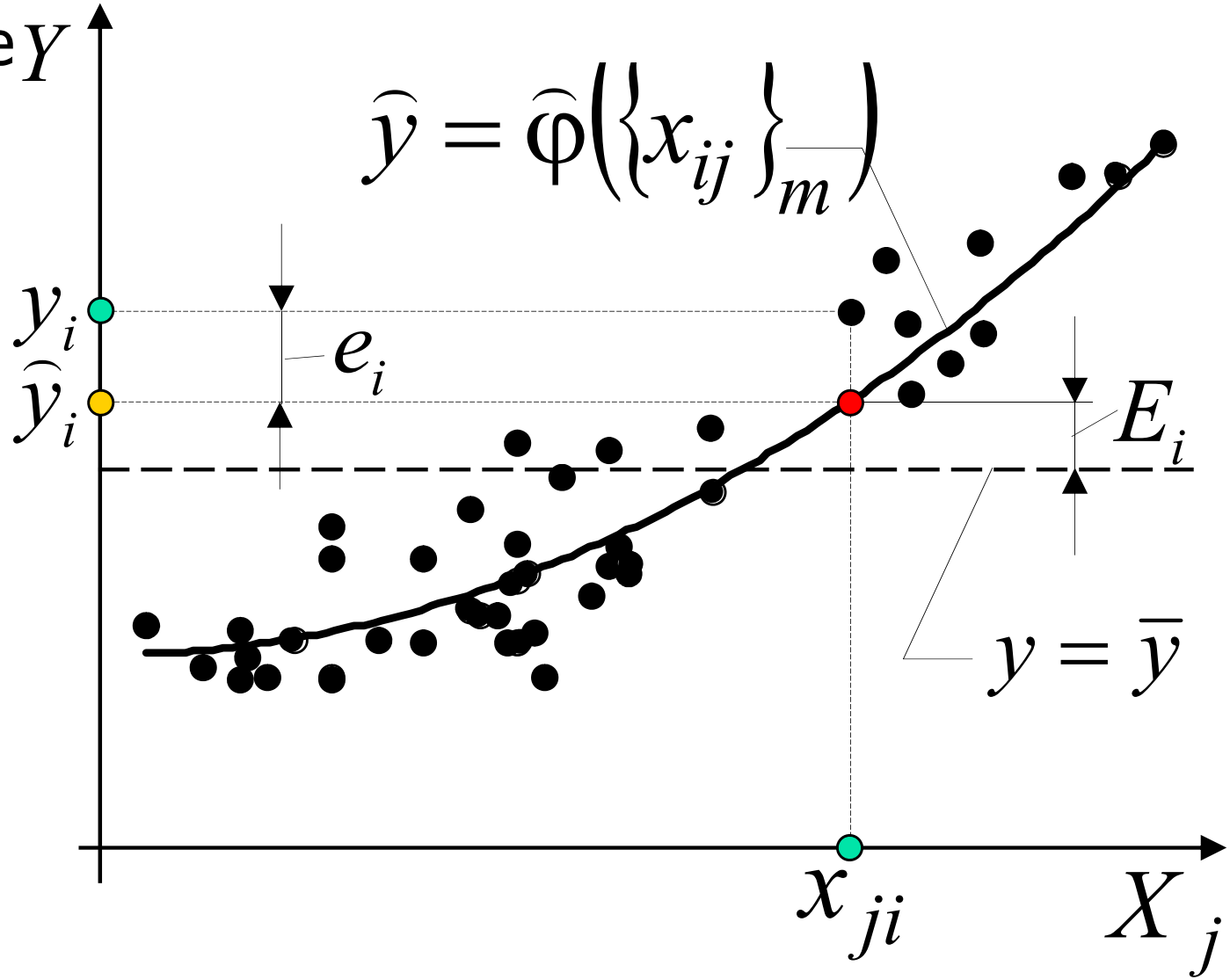
y_i - фактическое значение отклика

$\widehat{y}_i$ - 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

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

МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ АППРОКСИМАЦИЯ

ИСТИННАЯ ЗАВИСИМОСТЬ

$$y = \beta_0 + \beta_1 X_1 + \ldots + \beta_j X_j + \ldots \beta_m X_m$$

ИСТИННЫЕ КОЭФФИЦИЕНТЫ

УРАВНЕНИЕ РЕГРЕССИИ

$$\hat{y} = b_0 + b_1 X_1 + \ldots + b_j X_j + \ldots b_m X_m$$

КОЭФФИЦИЕНТЫ РЕГРЕССИИ

$$b_j \approx \beta_j$$

ЛИНЕЙН(<Y>; <X>; К1; К2),

ПРОБЛЕМА ЗНАЧИМОСТИ КОЭФФИЦИЕНТОВ РЕГРЕССИИ

Коэффициенты регрессии b_j являются случайными величинами с математическими ожиданиями β_j и дисперсиями $D(\beta_j)$, которым соответствуют стандартные отклонения S_{bj}.

УСЛОВИЕ ЗНАЧИМОСТИ КОЭФФИЦИЕНТОВ РЕГРЕССИИ

Коэффициент регрессии признается статистически значимым, если выполняется условие:

$$\underbrace{t_{bj} = \frac{|b_j|}{S_{bj}}}_{\text{Расчетное число Стьюдента}} > \underbrace{t[\alpha; n-k]}_{\text{Табличное число Стьюдента}}$$

СТЬЮДРАСПОБР(α;n-k)

- Если условие значимости b_j не выполнено, то следует признать, что влияние фактора X_j на отклик несущественное.
- Рекомендуется повторить регрессионный анализ без учета фактора X_j.

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

ПОКАЗАТЕЛЬ ДОСТОВЕРНОСТИ АППРОКСИМАЦИИ

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} E_i^2} \frac{n-1}{n-k}$$

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

ТЕСТ ФИШЕРА

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

АНАЛИЗ ОСТАТКОВ

Остатки - отклонения действительных значений отклика от рассчитанных по уравнению регрессии.

Для i-го наблюдения $$e_i = y_i - \widehat{y}_i$$

Отклонения обусловлены наличием случайной составляющей ε_i,

ПРЕДПОЛОЖЕНИЯ ОТНОСИТЕЛЬНО СЛУЧАЙНОЙ СОСТАВЛЯЮЩЕЙ

- Это нормально распределенная случайная переменная.
- Математическое ожидание случайной составляющей равно нулю
- Дсперсия случайной составляющей постоянна
- В различных наблюдениях значения случайных составляющих не зависят друг от друга.

Для проверки гипотезы о независимости отклонений в различных наблюдениях оценивают автокорреляцию остатков с применением критерия Дарбина-Уотсона

Строгое условие отсутствия автокорреляции DW=2.

С учетом особенностей распределения критерия Дарбина-Уотсона, ориентировочно можно считать, что автокорреляция остатков отсутствует при DW=1,2-2,8.

В противном случае следует признать, что гипотеза о независимости остатков в рассматриваемом случае не верна.

$$DW = \frac{\sum_{i=2}^{n} \left(e_i - e_{i-1}\right)^2}{\sum_{i=1}^{n} e_i^2}$$

- Если анализ остатков обнаруживает несоответствия указанным гипотезам, то уравнение регрессии, относительно которого данные остатки получены, следует считать неудовлетворительным, т. к. правомерность применения МНК и указанных выше оценок и для множественного регрессионного анализа может быть поставлена под сомнение.
- В таком случае рекомендуют рассмотреть уравнение иного вида (например, нелинейное вместо линейного), включить неучтенные ранее факторы, выделить в области варьирования факторов различные подобласти.

ПОСТРОЕНИЕ МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ АППРОКСИМАЦИИ В MS EXCEL С ПРИМЕНЕНИЕМ ИНСТРУМЕНТА «РЕГРЕССИЯ»

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

Регрессия

Входные данные

Входной интервал Y: $A$2:$A$30

Входной интервал X: $B$2:$C$30

☑ Метки

☐ Константа - ноль

☑ Уровень надежности: 95 %

OK

Отмена

Справка

Параметры вывода

◉ Выходной интервал: $E$1

○ Новый рабочий лист:

○ Новая рабочая книга

Остатки

☑ Остатки

☐ График остатков

☐ Стандартизованные остатки

☐ График подбора

Нормальная вероятность

☐ График нормальной вероятности

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	ИСХОДНЫЕ ДАННЫЕ				ВЫВОД ИТОГОВ			УРАВНЕНИЕ РЕГРЕССИИ						
2	Sт, МПа	tкп,С	tсм,С					Sт = 1152,965 - 0,498*tкп -0,695*tсм						
3	364,0	788	560		*Регрессионная статистика*			ЗНАЧИМОСТЬ КОЭФФИЦИЕНТОВ						
4	341,2	788	584		Множественный R	0,937		t[0,05;n-k]	b0	Значим				
5	338,4	788	611		R-квадрат	0,879		2,0595	b1	Значим				
6	284,1	788	639		Нормированный R-квадрат	0,869			b2	Значим		АВТОКОРРЕЛЯЦИЯ ОСТАТКОВ		
7	330,9	788	657		Стандартная ошибка	18,56		НАДЕЖНОСТЬ АППРОКСИМАЦИИ						
8	281,9	788	715		Наблюдения	28		F[0,05;k-1;n-k]	Аппроксимация			DW	2,3644	
9	265,9	788	728					3,3852	надежная			Отсутствует		
10	336,4	834	560		Дисперсионный анализ									
11	320,5	834	584			*df*	*SS*	*MS*	*F*	*Значимость F*				
12	336,8	834	611		Регрессия	2	62268,036	31134,018	90,4187	3,6E-12				
13	299,4	834	639		Остаток	25	8608,291	344,332						
14	269,3	834	657		Итого	27	70876,327							
15	240,3	834	715											
16	206,3	834	728			*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 90,0%*	*Верхние 90,0%*	
17	334,2	875	560		Y-пересечение	1152,695	73,484	15,6863	1,9E-14	1001,352	1304,039	1027,174	1278,217	
18	290,3	875	584		tкп,С	-0,498	0,073	-6,7941	4,03E-07	-0,649	-0,347	-0,623	-0,373	
19	318,5	875	611		tсм,С	-0,695	0,060	-11,6051	1,47E-11	-0,818	-0,572	-0,797	-0,593	
20	283,2	875	639											
21	281,0	875	657											
22	233,1	875	715											
23	200,0	875	728		ВЫВОД ОСТАТКА									
24	304,4	917	560											
25	307,5	917	584		*Наблюдение*	*Предсказанное Sт, МПа*	*Остатки*	$(e_i - e_{i-1})^2$						
26	268,8	917	611		1	371,3	-7,3	53,6						
27	225,4	917	639		2	354,6	-13,4	37,48						
28	270,3	917	657		3	335,9	2,5	254,80						

РЕЗУЛЬТАТ «РЕГРЕССИЯ»

ПРОВЕРКА ЗНАЧИМОСТИ КОЭФФИЦИЕНТОВ РЕГРЕССИИ

=СТЬЮДРАСПОБР(0,05;2;F8-3)

=ЕСЛИ(ABS(H17)>$H$5;"Значим";"Не значим")

=ЕСЛИ(ABS(H18)>$H$5;"Значим";"Не значим")

=ЕСЛИ(ABS(H19)>$H$5;"Значим";"Не значим")

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	ИСХОДНЫЕ ДАННЫЕ				ВЫВОД ИТОГОВ			УРАВНЕНИЕ РЕГРЕССИИ					
								Sт = 1152,965 - 0,498*tкп -0,695*tсм					
					...ая статистика			ЗНАЧИМОСТЬ КОЭФФИЦИЕНТОВ					
4	341,2	788	584		Множественный R	0,937		t[0,05;n-k]	b0	Значим			
5	338,4	788	611		R-квадрат	0,879		2,0595	b1	Значим			
									b2	Значим		АВТОКОРРЕЛЯЦИЯ ОСТАТКОВ	
								НАДЕЖНОСТЬ АППРОКСИМАЦИИ					
								F[0,05;k-1;n-k]	Аппроксимация			DW	2,3644
								3,3852	надежная			Отсутствует	
11	320,5	834	584					MS	F	Значимость F			
								31134,018	90,4187	3,6E-12			
								344,332					
15	240,3	834	715										
16	206,3	834	728			Коэффициенты	Стандартная ошибка	t-статистика	P-Значение	Нижние 95%	Верхние 95%	Нижние 90,0%	Верхние 90,0%
17	334,2	875	560		Y-пересечение	1152,695	73,484	15,6863	1,9E-14	1001,352	1304,039	1027,174	1278,217
18	290,3	875	584		tкп,C	-0,498	0,073	-6,7941	4,03E-07	-0,649	-0,347	-0,623	-0,373
19	318,5	875	611		tсм,C	-0,695	0,060	-11,6051	1,47E-11	-0,818	-0,572	-0,797	-0,593
20	283,2	875	639										
21	281,0	875	657										
22	233,1	875	715										
23	200,0	875	728		ВЫВОД ОСТАТКА								
24	304,4	917	560										
25	307,5	917	584		Наблюдение	Предсказанное Sт, МПа	Остатки	$(e_i - e_{i-1})^2$					
26	268,8	917	611		1	371,3	-7,3	53,6					
27	225,4	917	639		2	354,6	-13,4	37,48					
28	270,3	917	657		3	335,9	2,5	254,80					

ПРОВЕРКА СТАТИСТИЧЕСКОЙ НАДЕЖНОСТИ АППРОКСИМАЦИИ

=FРАСПОБР(0,05;2;F8-3)

=ЕСЛИ(H9>I12;”надежная”;”не надежная”)

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	ИСХОДНЫЕ ДАННЫЕ				ВЫВОД ИТОГОВ			УРАВНЕНИЕ РЕГРЕССИИ					
								Sт = 1152,965 - 0,498*tкп -0,695*tсм					
					ая статистика			ЗНАЧИМОСТЬ КОЭФФИЦИЕНТОВ					
4	341,2	788	584		Множественный R	0,937		t[0,05;n-k]	b0	Значим			
5	338,4	788	611		R-квадрат	0,879		2,0595	b1	Значим			
6	284,1	788	639		Нормированный R-квадрат	0,869			b2	Значим		АВТОКОРРЕЛЯЦИЯ ОСТАТКОВ	
7	330,9	788	657		Стандартная ошибка	18,56		НАДЕЖНОСТЬ АППРОКСИМАЦИИ					
8	281,9	788	715		Наблюдения	28		F[0,05;k-1;n-k]	Аппроксимация			DW	2,3644
9	265,9	788	728					3,3852	надежная			Отсутствует	
10	336,4	834	560		Дисперсионный анализ								
11	320,5	834	584			df	SS	MS	F	Значимость F			
12	336,8	834	611		Регрессия	2	62268,036	31134,018	90,4187	3,6E-12			
13	299,4	834	639		Остаток	25	8608,291	344,332					
14	260,2	834	657		Итого	27	70876,327						
15													
16	206,3	834	728			ициенты	артная ошибка	t-статистика	P-Значение	Нижние 95%	Верхние 95%	Нижние 90,0%	Верхние 90,0%
17	334,2	875	560		Y-пересечение	1152,695	73,484	15,6863	1,9E-14	1001,352	1304,039	1027,174	1278,217
18	290,3	875	584		tкп,С	-0,498	0,073	-6,7941	4,03E-07	-0,649	-0,347	-0,623	-0,373
19	318,5	875	611		tсм,С	-0,695	0,060	-11,6051	1,47E-11	-0,818	-0,572	-0,797	-0,593
20	283,2	875	639										
21	281,0	875	657										
22	233,1	875	715										
23	200,0	875	728		ВЫВОД ОСТАТКА								
24	304,4	917	560										
25	307,5	917	584		Наблюдение	Предсказанное Sm, МПа	Остатки	$(e_i - e_{i-1})^2$					
26	268,8	917	611		1	371,3	-7,3	53,6					
27	225,4	917	639		2	354,6	-13,4	37,48					
28	270,3	917	657		3	335,9	2,5	254,80					

ПРОВЕРКА АВТОКОРРЕЛЯЦИИ ОСТАТКОВ

=СУММ(H27:H53)/G13

=ЕСЛИ(M8<1,2;"Существует";ЕСЛИ(M8<2,8;"Отсутствует";"Существует"))

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	ИСХОДНЫЕ ДАННЫЕ				ВЫВОД ИТОГОВ			УРАВНЕНИЕ РЕГРЕССИИ					
								Sт = 1152,965 - 0,498*tкп -0,695*tсм					
						стика		ЗНАЧИМОСТЬ КОЭФФИЦИЕНТОВ					
						0,937		t[0,05;n-k]	b0	Значим			
5	338,4	788	611		R-квадрат	0,879		2,0595	b1	Значим			
6	284,1	788	639		Нормированный R-квадрат	0,869			b2	Значим		АВТОКОРРЕЛЯЦИЯ	
7	330,9	788	657		Стандартная ошибка	18,56		НАДЕЖНОСТЬ АППРОКСИМАЦИИ				ОСТАТКОВ	
8	281,9	788	715		Наблюдения	28		F[0,05;k-1;n-k]	Аппроксимация			DW	2,3644
9	265,9	788	728					3,3852	надежная			Отсутствует	
10	336,4	834	560		Дисперсионный анализ								
11	320,5	834	584			df	SS	MS	F	Значимость F			
12	336,8	834	611		Регрессия	2	62268,036	31134,018	90,4187	3,6E-12			
13	299,4	834	639		Остаток	25	8608,291	344,332					
14	269,3	834	657		Итого	27	70876,327						
16	206,3	834	728			ициенты	ошибка	статистика	Значение	95%	95%	90,0%	Верхние 90,0%
17	334,2	875	560		Y-пересечение	1152,695	73,484	15,6863	1,9E-14	1001,352	1304,039	1027,174	1278,217
18	290,3	875	584		tкп,С	-0,498	0,073	-6,7941	4,03E-07	-0,649	-0,347	-0,623	-0,373
19	318,5	875	611		tсм,С	-0,695	0,060	-11,6051	1,47E-11	-0,818	-0,572	-0,797	-0,593
20	283,2	875	639										
21	291,0	875	657										
22	233,1	875	715										
23	200,0	875	728		ВЫВОД ОСТАТКА								
24	304,4	917	560										
25	307,5	917	584		Наблюдение	Предсказанное Sт, МПа	Остатки	$(e_i - e_{i-1})^2$					
26	268,8	917	611		1	371,3	-7,3	53,6					
27	225,4	917	639		2	354,6	-13,4	37,48					
28	270,3	917	657		3	335,9	2,5	254,80					

Связь между какими величинами анализировалась?

Анализировалась связь между пределом текучести металла σ_T, температурой конца прокатки tкп и смотки tсм на ШСГП.

Значимы ли коэффициенты регрессии?

С доверительной вероятностью p = 95% коэффициенты регрессии b(tкп)=-0,498 и b(tсм)=-0,695 являются статистически значимыми, т. к. соответствующие числа Стьюдента |t(tкп)|=6,7941 и |t(tсм)|=11,6051 больше табличного t[0,05;25]=2,0595

Как выглядит уравнение множественной регрессии ?

Множественная линейная аппроксимация зависимости предела текучести металла от температуры конца прокатки и смотки на ШСГП имеет вид:

$$\sigma_T = 1152{,}695 - 0{,}498 t_{кп} - 0{,}695 t_{см}$$

Можно ли считать полученное уравнение множественной регрессии статистически надежной аппроксимацией анализируемой зависимости?

С доверительной вероятностью 95% полученное уравнение регрессии можно считать статистически надежной аппроксимацией исследуемой зависимости, т. к. рассчитанное число Фишера F=90,4187 больше табличного F[0,05;2;25]=3,3352