- **Регрессия в теории принятия решений**

Простая регрессия представляет собой регрессию между двумя переменными — **y** и **x**, т. е. модель вида:

$$y = f(x),$$

где:

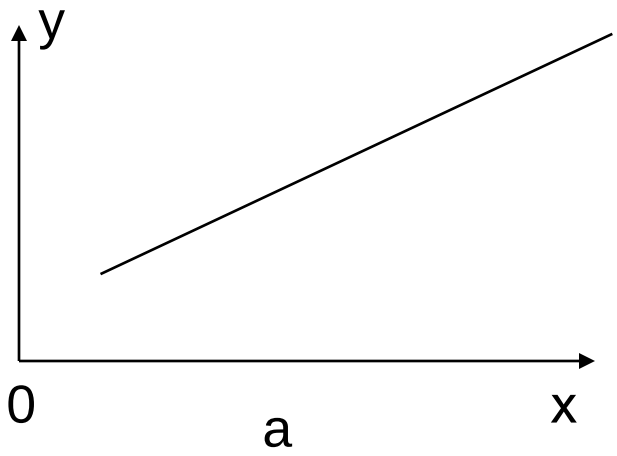
y – зависимая переменная (результативный признак);

x – независимая, или объясняющая, переменная (признак-фактор).

Основные типы кривых, используемые при количественной оценке связей между двумя переменными

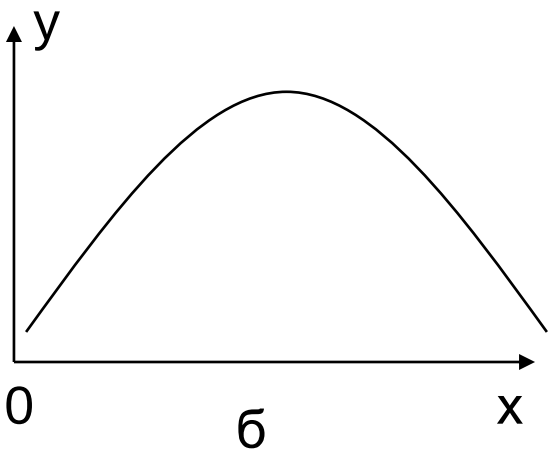
$а) \hat{y}_x = a + b \cdot x;$

линейная регрессия



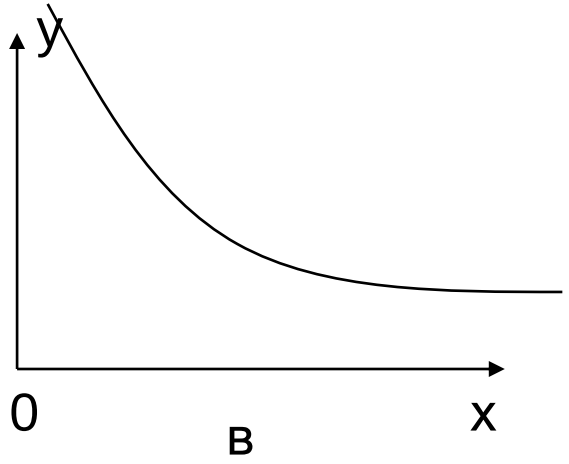
$б) \hat{y}_x = a + b \cdot x + c \cdot x^2$

полином второй степени



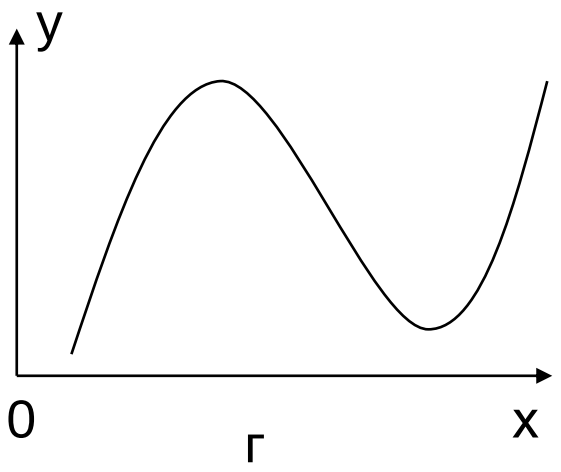
$в) \hat{y}_x = a + b / x;$

равносторонняя гипербола



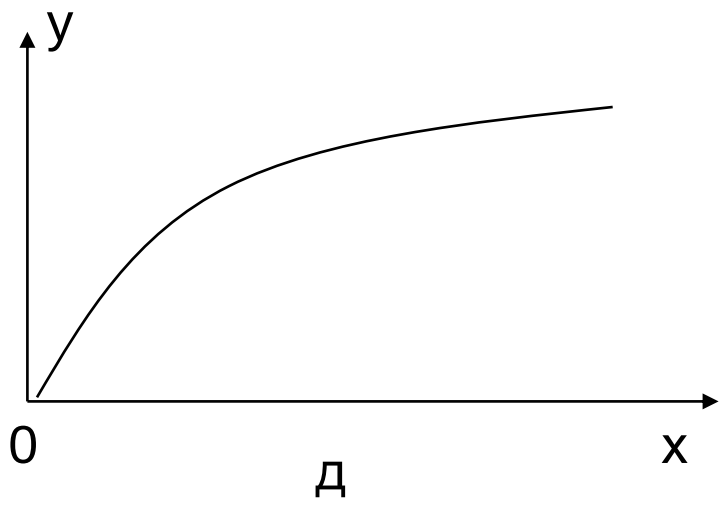
$г) \hat{y}_x = a + b \cdot x + c \cdot x^2 + d \cdot x^3$

полином третьей степени



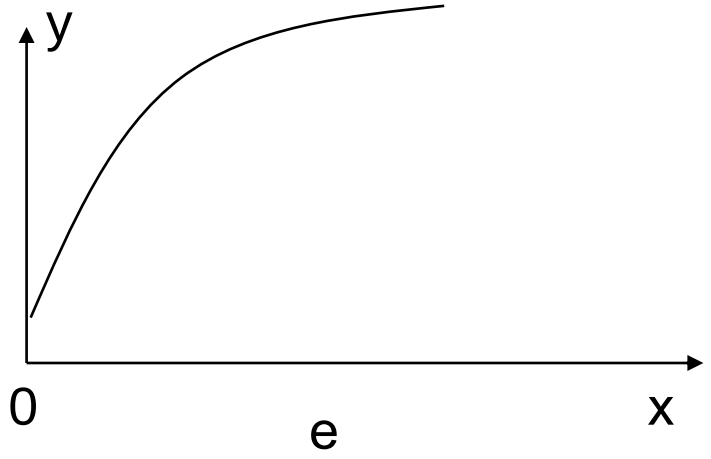
$д) \hat{y}_x = a \cdot x^b,$

степенная



$е) \hat{y}_x = a \cdot b^x,$

показательная

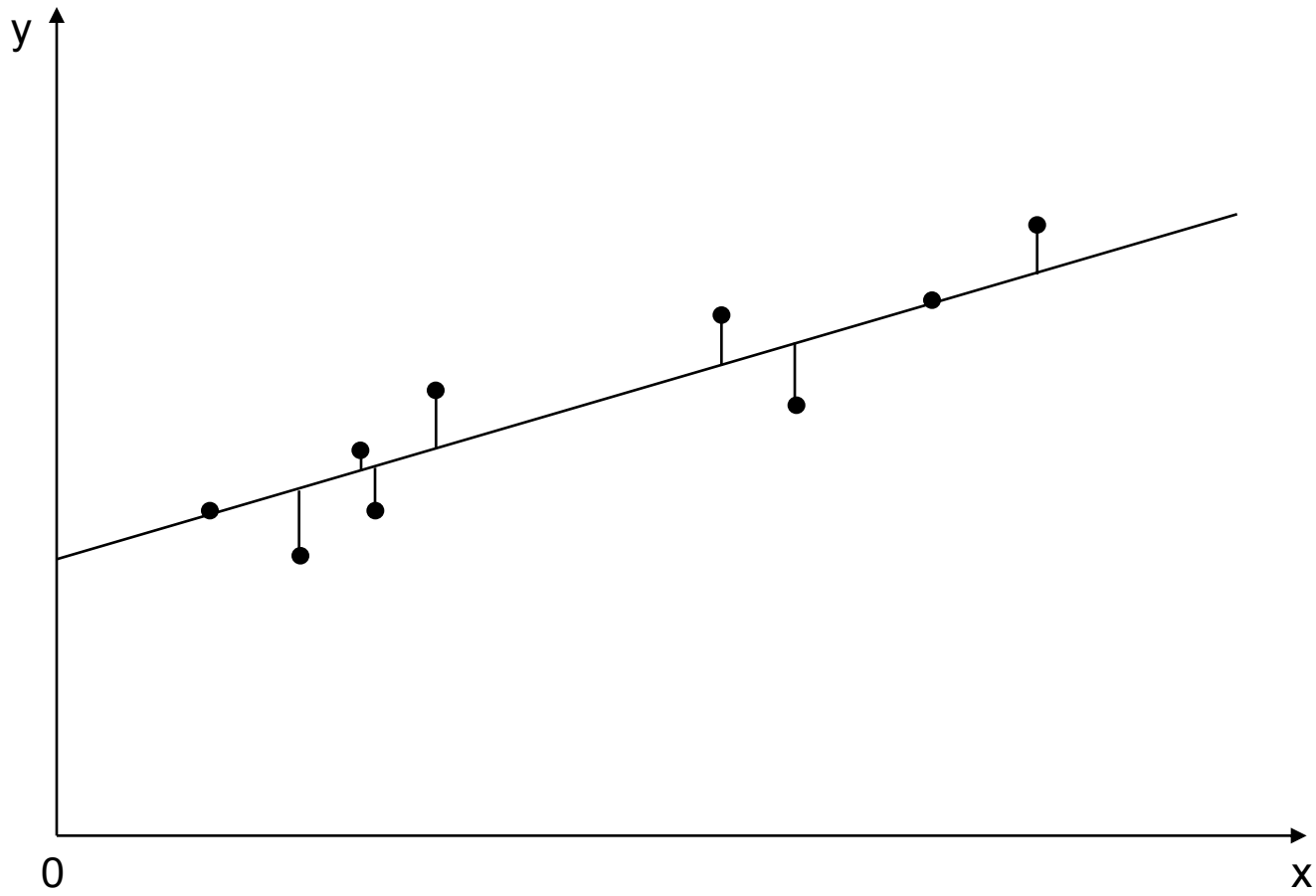


Классический подход к оцениванию параметров линейной регрессии основан на **методе наименьших квадратов** (МНК).

МНК позволяет получить такие оценки параметров *a* и *b*, при которых сумма квадратов отклонений фактических значений результативного признака (y_i) от расчетных $\hat{y}_x$ (теоретических) минимальна:

$$\sum_i \left(y_i - \hat{y}_{x_i}\right)^2 \to \min$$

- Геометрический смысл МНК: из всего множества линий линия регрессии на графике выбирается так, чтобы сумма квадратов расстояний по вертикали между точками и этой линией была бы минимальной



Обозначим $\varepsilon_i = y_i - \hat{y}_{x_i}$,

$$S = \sum_{i=1}^{n} \varepsilon_i^{\,2} \to \min$$

$$S = \sum_i \left(y_i - \hat{y}_{x_i} \right)^2 = \sum_i \left(y - a - b \cdot x \right)^2$$

$$(1)\quad \begin{cases} \dfrac{dS}{da} = -2\sum_{i=1}^{n} y_i + 2\cdot n\cdot a + 2\cdot b\sum_{i=1}^{n} x_i = 0; \\ \dfrac{dS}{db} = -2\sum_{i=1}^{n} y_i x_i + 2\cdot a\sum_{i=1}^{n} x_i + 2\cdot b\sum_{i=1}^{n} x_i^2 = 0. \end{cases}$$

для оценки параметров a и b получим следующую систему нормальных уравнений

$$\begin{cases} n \cdot a + b \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^{\,2} = \sum_{i=1}^{n} x_i \cdot y_i \end{cases}$$

Формулы расчета параметров a и b:

$$a = \bar{y} - b \cdot \bar{x}$$

$$b = \frac{\overline{yx} - \bar{y} \cdot \bar{x}}{\overline{x^2} - \bar{x}^2}$$

b - коэффициент регрессии. Его величина показывает среднее изменение результата с изменением фактора на одну единицу.

$$\sigma_x^2 = \overline{x^2} - \bar{x}^2$$

Линейный коэффициент корреляции является показателем тесноты связи:

$$r_{xy} = b\frac{\sigma_x}{\sigma_y} = \frac{\overline{yx} - \overline{y}\cdot\overline{x}}{\sigma_x\sigma_y}$$

Линейный коэффициент корреляции должен находится в границах:

$$-1 \le r_{xy} \le 1$$

Для характеристики силы связи можно использовать шкалу *Чеддока.*

Показатель тесноты связи	**0,1 – 0,3**	**0,3 – 0,5**	**0,5 – 0,7**	**0,7 – 0,9**	**0,9 – 0,99**
Характеристика силы связи	**Слабая**	**Умеренная**	**Заметная**	**Высокая**	**Весьма высокая**

- **Коэффициент детерминации** r_{yx}^2 характеризует долю дисперсии результативного признака :

- Величина 1- r^2 характеризует долю дисперсии *y,* вызванную влиянием остальных не учтенных в модели факторов.

- **Пример.** Предположим по группе предприятий, выпускающих один и тот же вид продукции, рассматривается зависимость затрат на производство(y) от выпуска продукции(x)

Выпуск продукции, тыс. ед. *(x)*	Затраты на производство, млн руб. *(y)*
1	30
2	70
4	150
3	100
5	170
3	100
4	150

- Система нормальных уравнений будет иметь вид

$$\begin{cases} 7a + 22b = 770 \\ 22a + 80b = 2820 \end{cases}$$

- a = -5,798, b= 36,8443,
- $r^2 = 0{,}982$.
- уравнение регрессии:

$$\hat{y}_x = -5{,}79 + 36{,}84x$$

Оценка существенности уравнения линейной регрессии.

- ***F критерий Фишера*** - оценивает качество уравнения регрессии - состоит в проверке гипотезы H_0 (о том, что коэффициент регрессии равен нулю, т.е. $b = 0$, т.е. фактор x не оказывает влияния на результат y).

$$F = \frac{D_{факт}}{D_{ост}}$$

- дисперсии на одну степень свободы

$$D_{общ} = \frac{\sum (y - \bar{y})^2}{n-1}$$

$$D_{факт} = \frac{\sum (\hat{y}_x - \bar{y})^2}{1}$$

$$D_{ост} = \frac{\sum (y - \hat{y}_x)^2}{n-2}$$

- Число степеней свободы остаточной суммы квадратов при линейной парной регрессии составляет $n - 2$,
- общей суммы квадратов – $n - 1$,
- для факторной суммы квадратов – 1,

Имеем равенство:

$$n - 1 = 1 + (n - 2).$$

$$\sum (\hat{y}_x - \bar{y})^2 = r^2 \cdot \sigma_y^2 \cdot n$$

$$\sum (y - \hat{y}_x)^2 = (1 - r^2) \cdot \sigma_y^2 \cdot n$$

$$F_{факт} = \frac{r^2}{1 - r^2}(n - 2)$$

- *n* - число наблюдений

- Значение *F*-критерия признается достоверным, если оно больше табличного. В этом случае гипотеза H_0 отклоняется.

•Если $F_{табл} < F_{факт}$, то H_o – гипотеза о случайной природе оцениваемых характеристик отклоняется и признается их статистическая значимость и надежность.

•Если $F_{табл} > F_{факт}$, то гипотеза H_o не отклоняется и признается статистическая незначимость и ненадежность уравнения регрессии.

- **Таблица значений F-критерия Фишера при уровне значимости α =0,05**

k1 k2	1	2	3	4	5	6	8	12	24	∞
1	161,45	199,50	215,72	224,57	230,17	233,97	238,89	243,91	249,04	254,32
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,45	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,64	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,77	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	3,84	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,12	2,93

- Для оценки статистической значимости коэффициентов регрессии и корреляции рассчитываются

t-критерий Стьюдента

каждого из показателей и

доверительные интервалы.

$$t_a = \frac{a}{m_a} \qquad t_b = \frac{b}{m_b} \qquad t_r = \frac{r}{m_r}$$

$$m_a = S_{ocm} \frac{\sqrt{\sum x^2}}{n\sigma_x} \qquad m_b = \frac{S_{ocm}}{\sigma_x \sqrt{n}} \qquad m_r = \sqrt{\frac{1-r^2}{n-2}}$$

- Если $t_{табл} < t_{факт}$ то гипотеза H_0 - о незначимости параметра отклоняется, т.е. соответствующие параметры не случайно отличаются от нуля и сформировались под влиянием систематически действующего фактора x.

- Если $t_{табл} > t_{факт}$ то гипотеза H_0 не отклоняется и признается случайная природа формирования соответствующих параметров уравнения регрессии .

- **Критические значения t-критерия Стьюдента при уровне значимости 0,10; 0,05;**

d.f.	a		
	0,10	0,05	0,01
1	6,3138	12,706	63,657
2	2.9200	4,3027	9,9248
3	2,3534	3,1825	5,8409
4	2,1318	2,7764	4,6041
5	2,0150	2,5706	4,0321
6	1,9432	2,4469	3,7074
7	1,8946	2,3646	3,4995

- ***доверительный интервал***
- для расчета доверительного интервала определяем ***предельную ошибку*** Δ

$$\Delta_b = t_{табл} m_b \qquad \Delta_a = t_{табл} m_a$$

- для коэффициентов регрессии границы доверительного интервала составят:

$$(a - \Delta_a, a + \Delta_a) \qquad (b - \Delta_b, b + \Delta_b)$$

Если в границы доверительного интервала попадает 0, то оцениваемый параметр принимается нулевым, так как он не может одновременно принимать и положительное, и отрицательное значения.

- Пример. Предположим по группе предприятий, выпускающих один и тот же вид продукции, рассматривается зависимость затрат на производство(y) от выпуска продукции(x)

Выпуск продукции, тыс. ед. *(x)*	Затраты на производство, млн руб. *(y)*
1	30
2	70
4	150
3	100
5	170
3	100
4	150

уравнение регрессии: $\hat{y}_x = -5{,}79 + 36{,}84x$

$r^2 = 0{,}982$, $r = 0{,}991$

$$F = \frac{0{,}982}{1 - 0{,}982} \cdot (7 - 2) = 273$$

$t_a = -0{,}78$ $t_b = 16{,}67$

- Доверительные интервалы
- $-22{,}39 \leq a \leq 10{,}801$
- $31{,}16 \leq b \leq 42{,}52$.

- ***Прогнозное значение*** y_p определяется путем подстановки в уравнение регрессии $\bar{y}_x = a + b \cdot x$ интересуемого (прогнозного) значения независимой переменной x_p .

пример

- Выполнить, по уравнению регрессии $y=280+5,6x$, прогноз заработной платы y при прогнозном значении среднедушевого прожиточного минимума x, составляющем 127% от среднего уровня (x=6700).

Нелинейная регрессия.

- Нелинейная регрессия определяется, как в линейной регрессии, методом наименьших квадратов (МНК).

в параболе второй степени ,

$$y = a_0 + a_1 \cdot x + a_2 \cdot x^2$$

заменяя переменные , $x = x_1 \quad x^2 = x_2$

получим двухфакторное уравнение линейной регрессии:

$$y = a_0 + a_1 \cdot x_1 + a_2 \cdot x_2$$

- для полинома k-го порядка

$$y = a_0 + a_1 \cdot x + a_2 \cdot x^2 + \ldots + a_k \cdot x^k$$

$$y = a_0 + a_1 \cdot x_1 + a_2 \cdot x_2 + \ldots + a_k \cdot x_k$$

- В уравнении равносторонней гиперболы –

$$y = a + \frac{b}{x} + \varepsilon$$

делаем замену $z=1/x$,

получаем линейное уравнение

$$y=a+bz$$

Для степенной модели $y = a \cdot x^b$

линеаризация производится путём логарифмирования обеих частей уравнения

$$\lg y = \lg a + b \lg x$$

с помощью замены

$$Y = \lg y, \quad X = \lg x, \quad A = \lg a$$

получаем линейное уравнение

$$Y = A + bX$$

Для показательной модели $y = a \cdot b^x$

линеаризация производится также с помощью логарифмирования обеих частей уравнения $\lg y = \lg a + x \lg b$

с помощью замены

$$Y = \lg y, \quad B = \lg b, \quad A = \lg a$$

получаем линейное уравнение

$$Y = A + B \cdot x$$

Пример.

- Линеаризовать модель:

$$\bar{y}_x = 50x^{0.6}$$

- Записать уравнение в явном виде

$ln y = 1{,}1 + 0{,}8 ln x$

- **Корреляция для нелинейной регрессии.**

$$\rho = \sqrt{1 - \frac{\Sigma(y - \hat{y}_x)^2}{\Sigma(y - \bar{y})^2}}$$

- Величина данного показателя находится в границах: $0 \le \rho \le 1$,

- При использовании линеаризуемых функций, затрагивающих преобразование зависимой переменной y, линейный коэффициент корреляции по преобразованным значениям признаков не совпадает с индексом корреляции

- проверка существенности в целом уравнения нелинейной регрессии осуществляется с помощью **F-критерия Фишера**

- оценка статистической значимости коэффициентов регрессии и корреляции осуществляется с помощью ***t-критерия Стьюдента*** и **доверительных интервалов**

- Наилучшая модель выбирается на основе показателей: корреляции, детерминации, F-критерия Фишера, t-критерия Стьюдента.

Пример.

- По территориям Центрального района РФ приводятся данные за 200Х год:
- y - выплаты социального характера, тыс. руб.
- x- прожиточный минимум в среднем на душу населения, тыс. руб.

Район	y	x
Брянская обл.	6,9	289
Владимировская обл.	8,7	334
Ивановская обл.	6,4	300
Калужская обл.	8.4	343
Костромская обл.	6,1	356
Орловская обл.	9,4	289
Рязанская обл.	11,0	341
Смоленская обл.	6,4	327
Тверская обл.	9,3	357
Тульская обл.	8,2	352
Ярославская обл.	8,6	381

- Рассчитайте параметры уравнений линейной и гиперболической парной регрессии;

- Выберите лучшее уравнение регрессии;

- Постройте прогноз средней заработной платы и выплат социального характера, если прогнозное значение фактора увеличится на 21% от его среднего уровня

	a	b	r	r^2	F
y=a+bx	-209,2	2,938	0,75	0,57	11,99
y=a+b/x	1636,94	-288997,8	0,76	0,58	12,36

- уравнение регрессии:

$$y = -209,2 + 2,938x$$

$$y = 1636,94 - \frac{288997,8}{x}$$

$$Fтабл = 5,12$$

	t критерий Стьюдента		**доверительный интервал**	
	a	b	a	b
y=a+bx	-0,795	3,464	(-804,1; 385,74)	(1,01 ; 4,84)
y=a+b/x	6,134	-3,515	(1033,25 ; 2240,64)	(-474990; -103005,4)

$$t_{табл} = 2,26$$

- Вывод: гиперболическая модель лучше описывает статистические данные. Прогноз необходимо делать по гиперболическому уравнению регрессии.

$$y = 1636,94 - \frac{288997,8}{x}$$

$$x_{ср} = 309,545$$

$$x_{р} = 309,545 \cdot 1,21 = 374,55$$

$$y_p = 1636,94 - \frac{288997,8}{374,55} = 865,357$$

- Ответ: При увеличении в Центральном районе РФ прожиточного минимума на 21% выплаты социального характера составят 865,357 тыс. руб.

Решение задач ЛП с помощью надстройки "Поиск решения"

- Пример. Решить задачу ЛП

$$f = 2,2x_1 + 1,95x_2 + 2,87x_3 \to \min$$

$$\begin{cases} 10x_1 + 6x_2 + 12x_3 \geq 50 \\ 7x_1 + 10x_2 + 11x_3 \geq 45 \\ x_1, x_2, x_3 \geq 0 \end{cases}$$

- 1. Математическую формулировку задачи необходимо оформить в виде таблицы, отражающей основные зависимости.

	A	B	C	D	E
1	коэффициенты матрицы ограничений			вектор ресурсов	произведение AX
2	10	6	12	50	0
3	7	10	11	45	0
4					
5	вектор решений				
6	x1	x2	x3		
7	0	0	0		
8					
9	коэффициенты целевой функции			Значение целевой функции	
10	2,2	1,95	2,87	0	

- E2 = СУММПРОИЗВ(A2:C2;A7:C7)
- E3 = СУММПРОИЗВ(A3:C3;A7:C7)
- E10 = СУММПРОИЗВ(A10:C10;A7:C7)

- 2. Вызов надстройки «Поиск решения»:

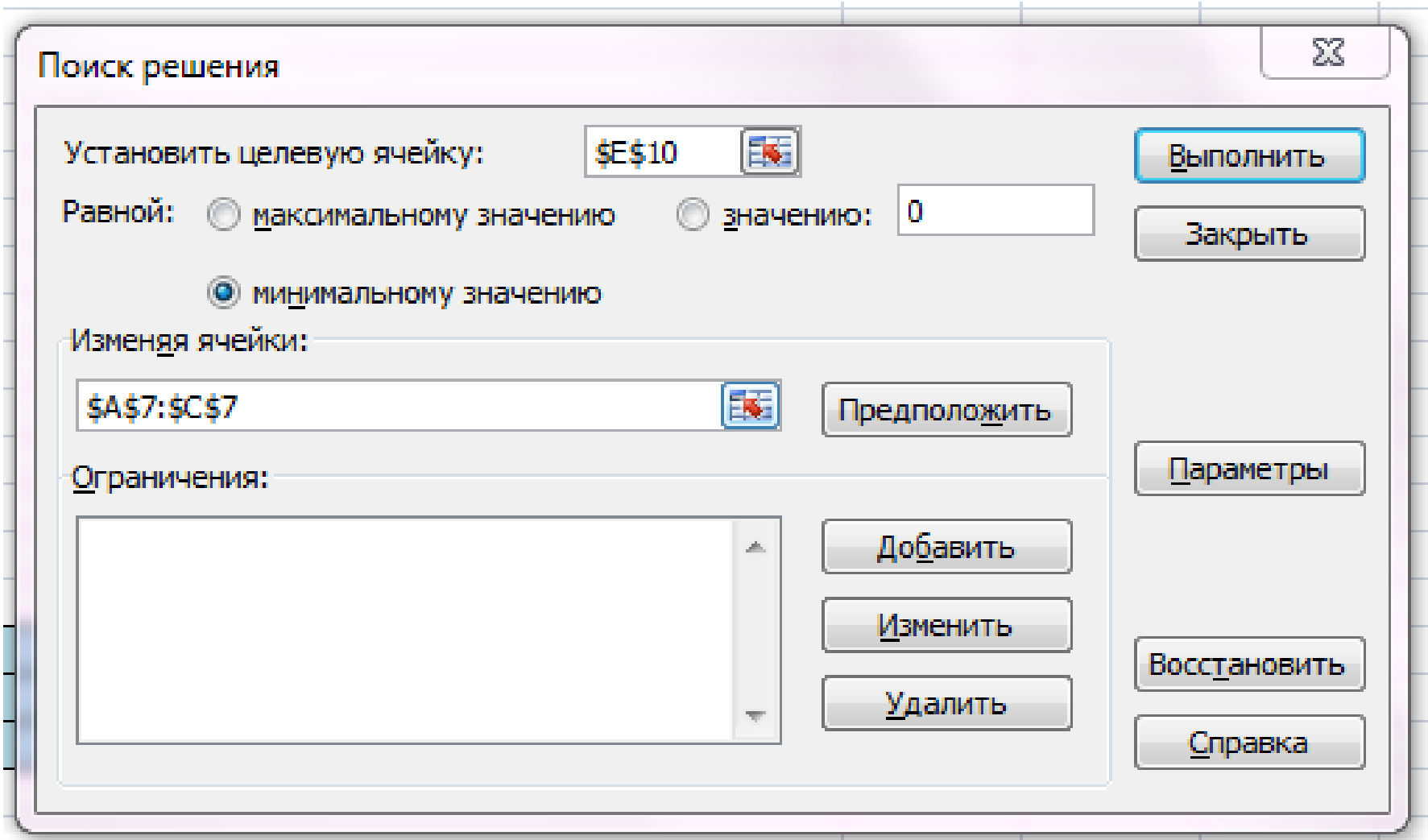
Данные - Поиск решения

Настройка пакета: параметры – надстройки – перейти – «v» поиск решения - ok

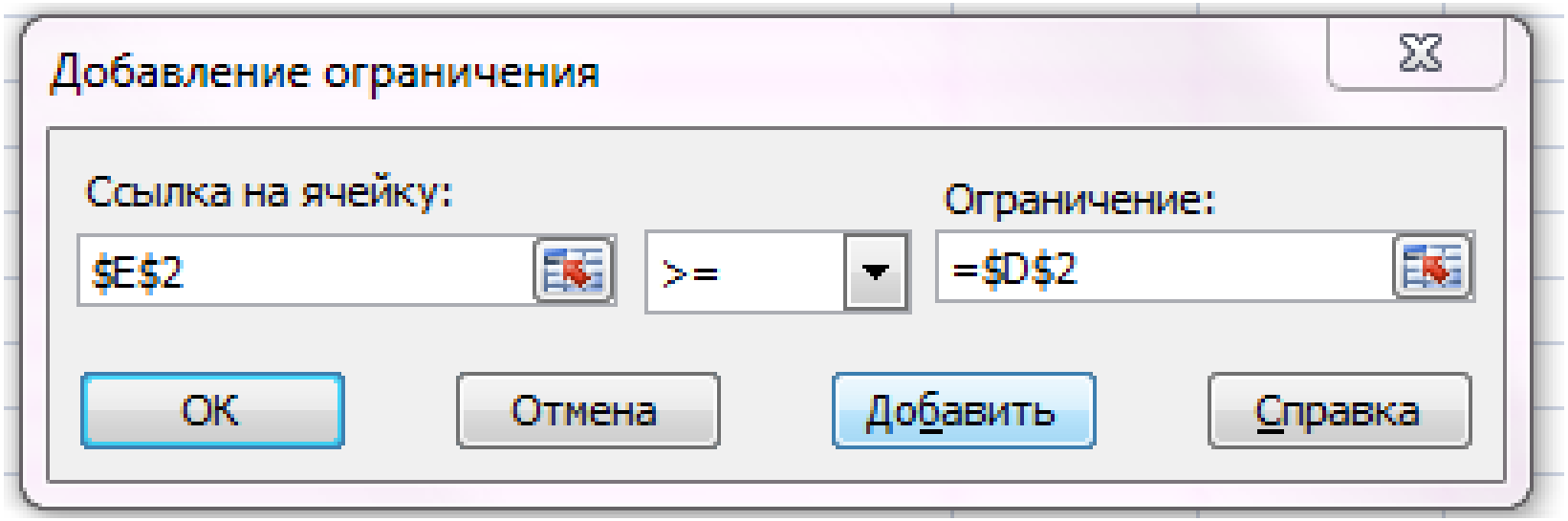
Целевая ячейка = значение целевой функции

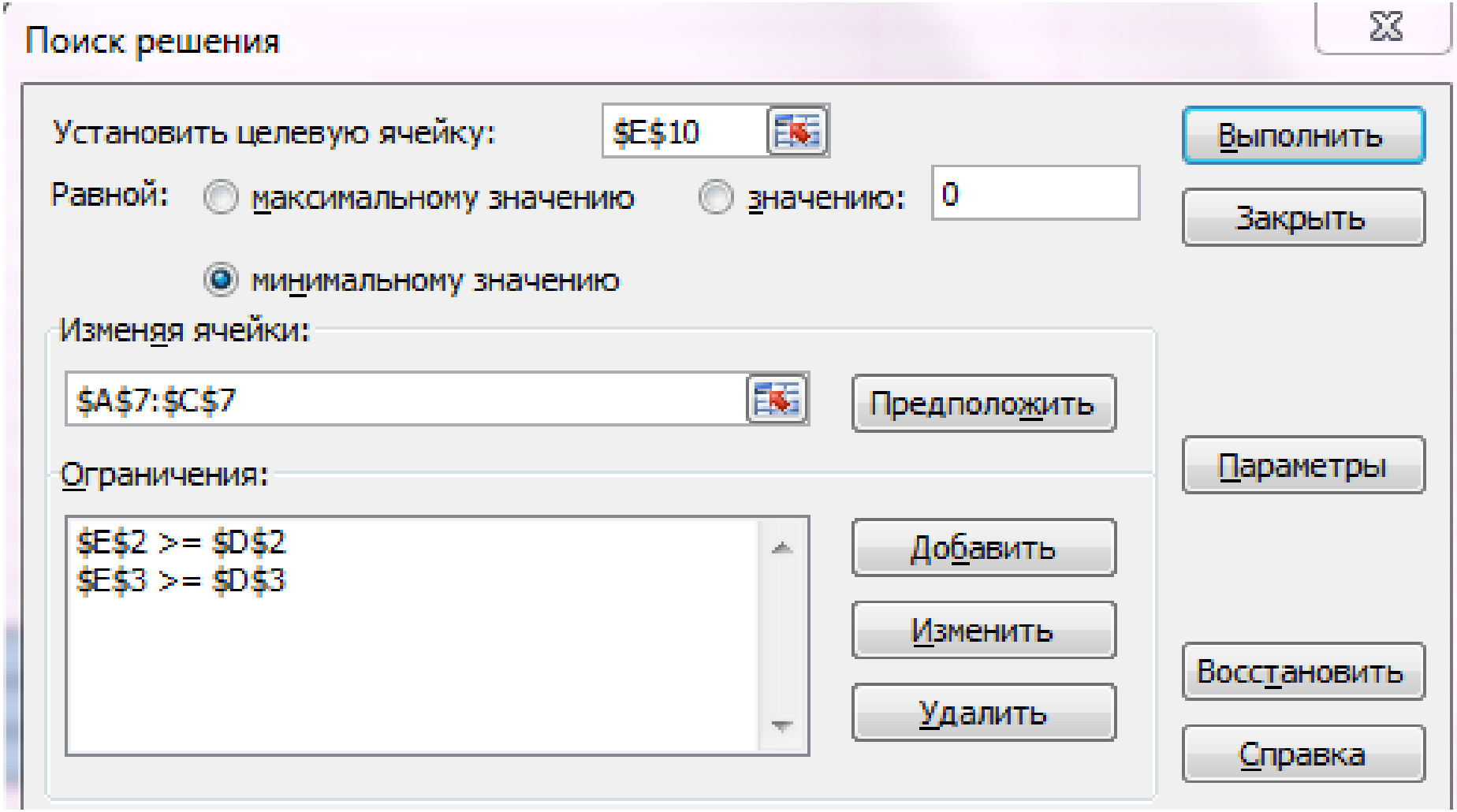
Изменяя ячейки = вектор решений

Равной (min,max)

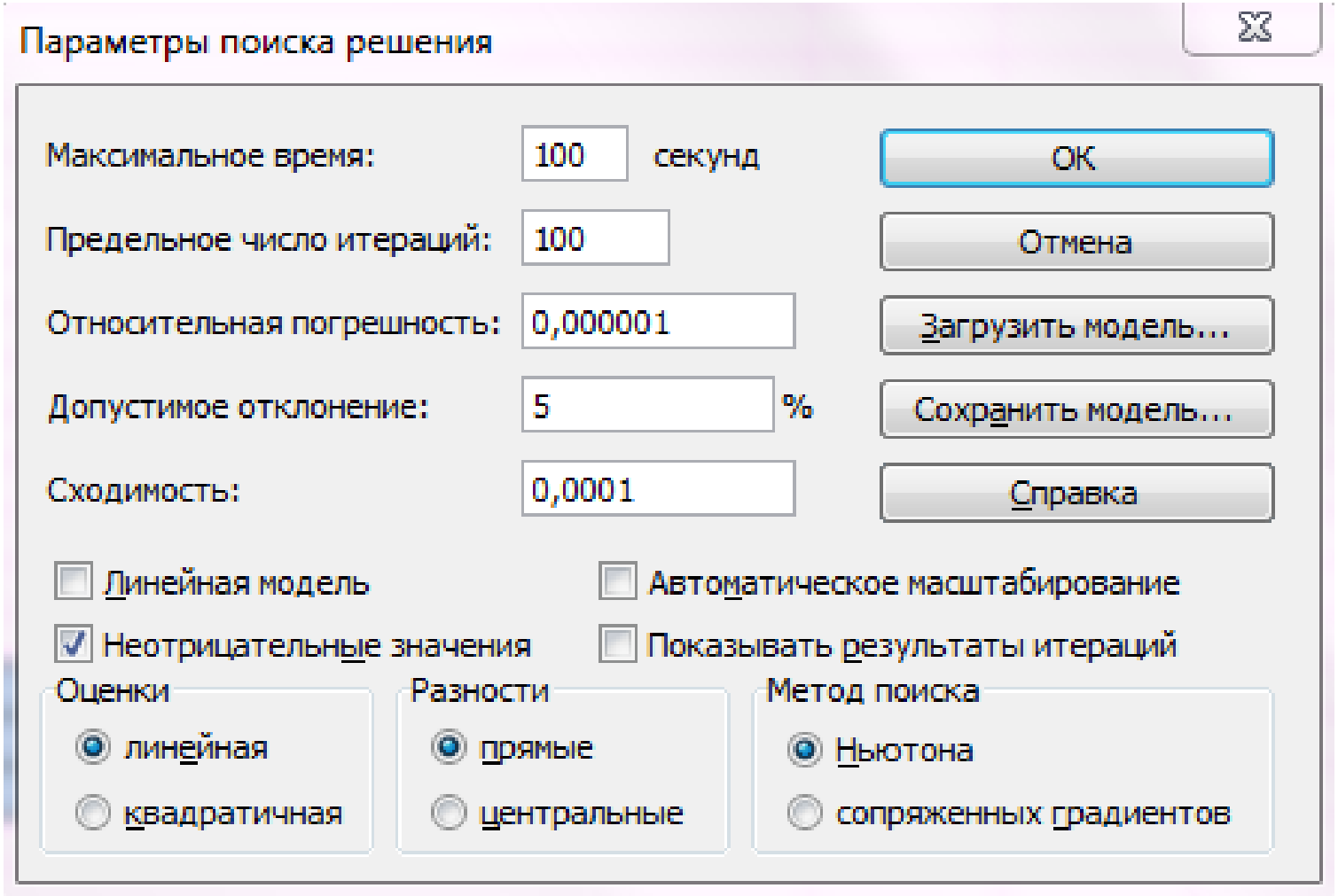


- Добавление ограничения
- Ссылка на ячейку = столбец AX
- Ограничение = вектор ресурсов

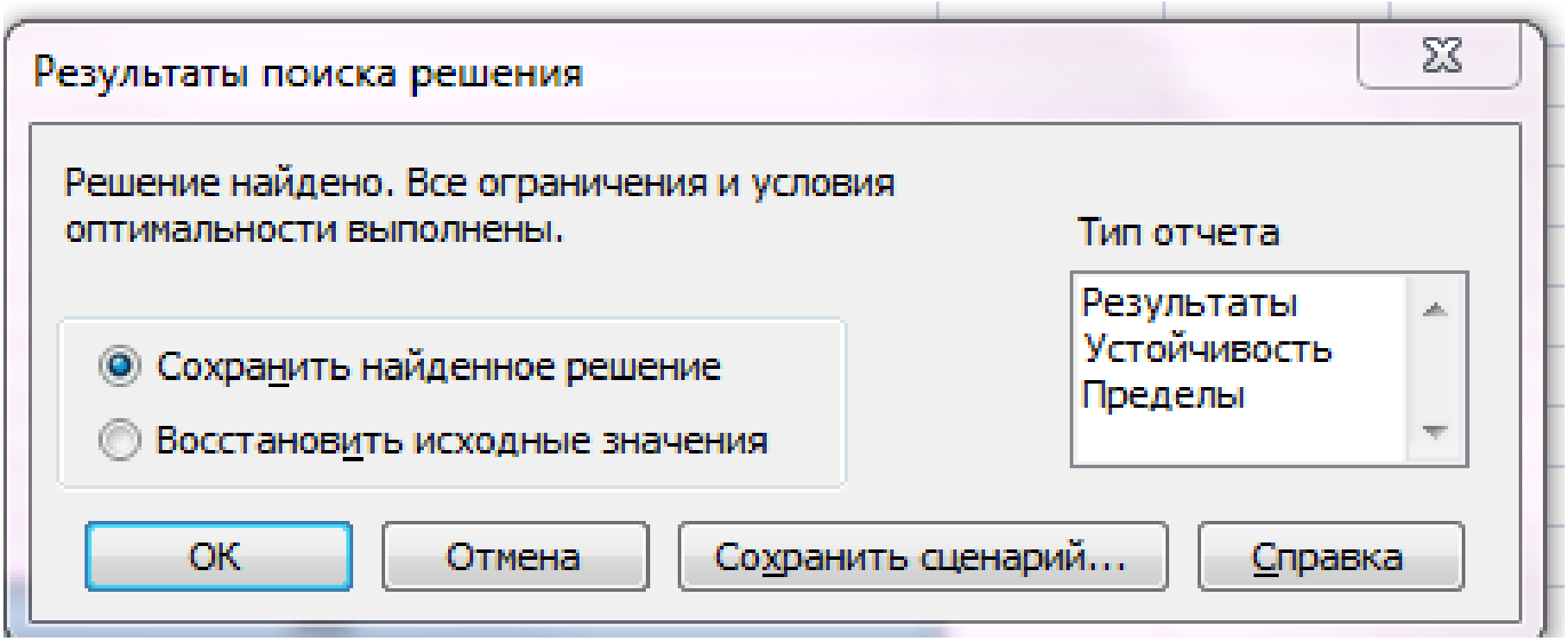



Поиск решения
Установить целевую ячейку:
$E$10
Выполнить
Равной:
максимальному значению
значению:
0
Закрыть
минимальному значению
Изменяя ячейки:
$A$7:$C$7
Предположить
Параметры
Ограничения:
$E$2 >= $D$2
$E$3 >= $D$3
Добавить
Изменить
Удалить
Восстановить
Справка

- Если $x_i \geq 0$, то $\longrightarrow$ параметры $\longrightarrow$ «v» неотрицательные значения $\longrightarrow$ ok



- 3. Выполнить


Результаты поиска решения
Решение найдено. Все ограничения и условия оптимальности выполнены.
Тип отчета
Результаты
Устойчивость
Пределы
Сохранить найденное решение
Восстановить исходные значения
ОК
Отмена
Сохранить сценарий...
Справка

	A	B	C	D	E
1	коэффициенты матрицы ограничений			вектор ресурсов	произведение AX
2	10	6	12	50	50
3	7	10	11	45	45
4					
5	вектор решений				
6	x1	x2	x3		
7	**0,38**	**0**	**3,85**		
8					
9	коэффициенты целевой функции			Значение целевой функции	
10	2,2	1,95	2,87	**11,88**	