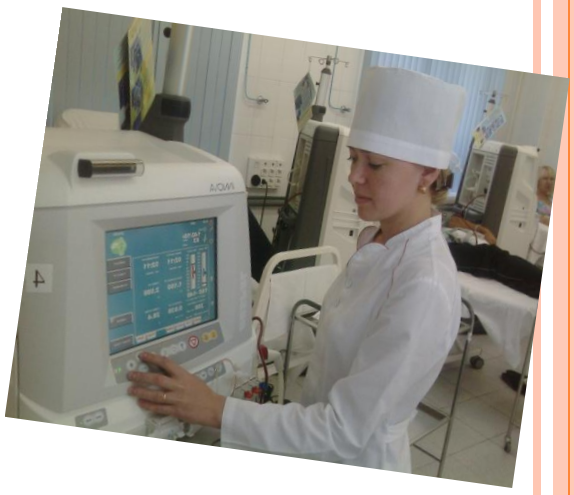


Лекция 9
Регрессионный в Excel.
Дисперсионный анализ.

Составитель: доц. Космачева И.М.

DATA MINING – КЛАССЫ РЕШАЕМЫХ ЗАДАЧ

- Классификационная и регрессионная модели

устанавливают закономерности между входными и выходными переменными.

- Если входные и выходные переменные модели непрерывные — перед нами задача **регрессии**.
- Если выходная переменная одна и она является **дискретной** (метка класса), то речь идет о задаче классификации.

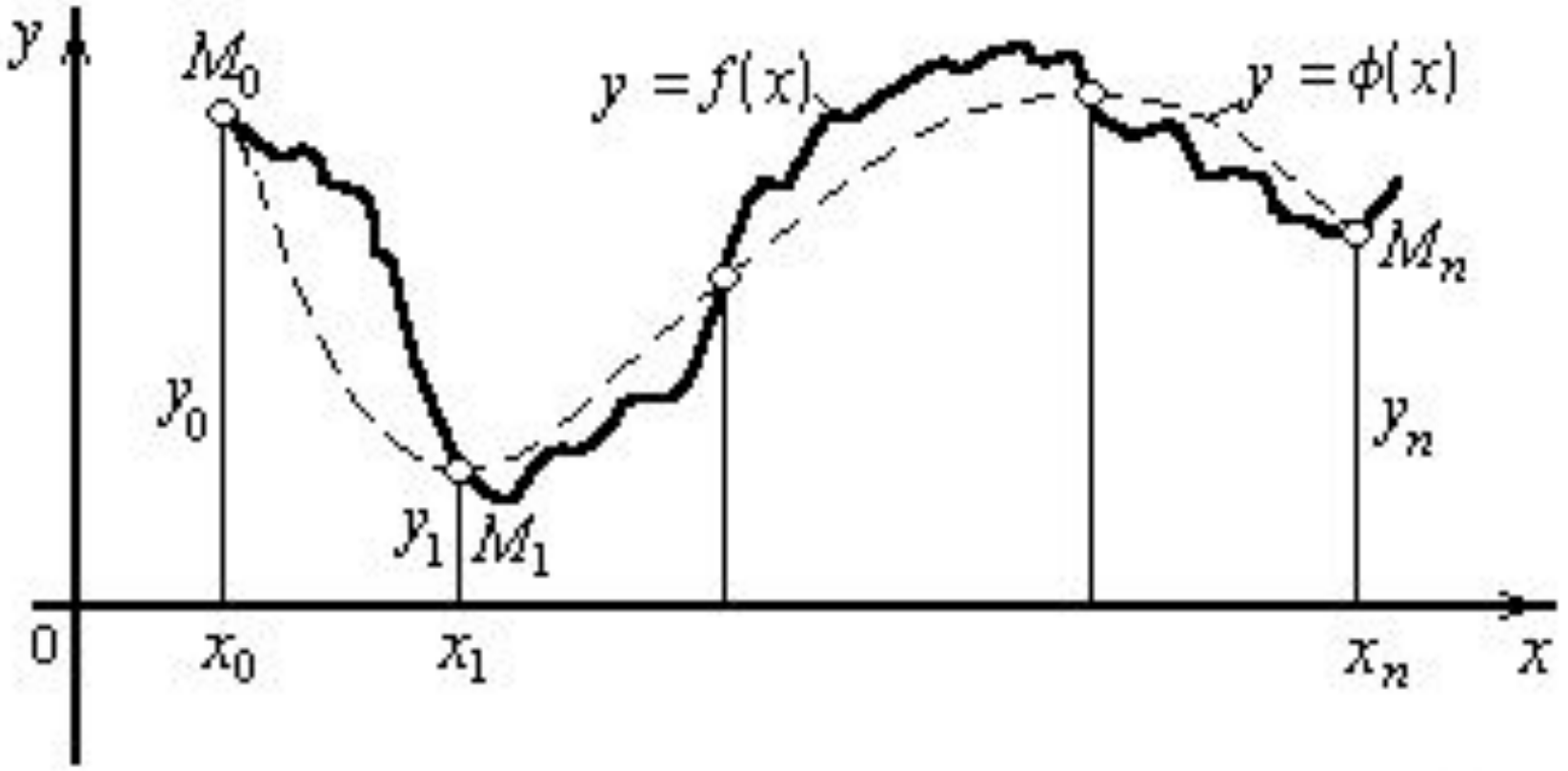
В *медицине* с помощью классификации и регрессии можно диагностировать заболевания на основе наблюдаемых симптомов (температура, давление, состав крови и т. д.), оценивать ожидаемые результаты лечения.

Регрессия

- **Цель регрессионного анализа** – по результатам наблюдений за входными и выходными величинами найти зависимость между входами и выходом, т.е. получить математическую модель.
- Нахождение функциональной зависимости между входными атрибутами и **непрерывным выходным** атрибутом.
- Задачи регрессионного анализа :

1. Прогнозирование ухудшения состояния пациента.
2. Оценка вероятности повторных рецидивов заболевания.
3. Расчет загруженности докторов при обслуживании населения.
4. Анализ влияния различных факторов на исследуемый.

Регрессия



Регрессия

- ***Регрессией Y на X называется функциональная зависимость между значениями x и соответствующими условными средними y(x).***
- Форма связи результативного признака Y с факторами X1, X2,…Xm называется **уравнением регрессии**. В зависимости от типа выбранного уравнения различают **линейную и нелинейную регрессию**, а в зависимости от количества факторов – **парную (простую, $m = 1$) и множественную (многофакторную, $m > 1$).**
- Регрессионный анализ связан с корреляционным (также часто встречается термин «корреляционно-регрессионный анализ».
- Корреляционный анализ позволяет сделать предположения о характере связи между изучаемыми факторами.

Нелинейная регрессия

На практике в качестве функции $f(x)$ для парной регрессии используются следующие виды функций:

Полиномиальная k-го порядка –

$$f(x) = \beta_0 + \beta_1 x + \ldots + \beta_k x^k.$$

Экспоненциальная – $f(x) = \beta_0 \exp(\beta_1 x)$.

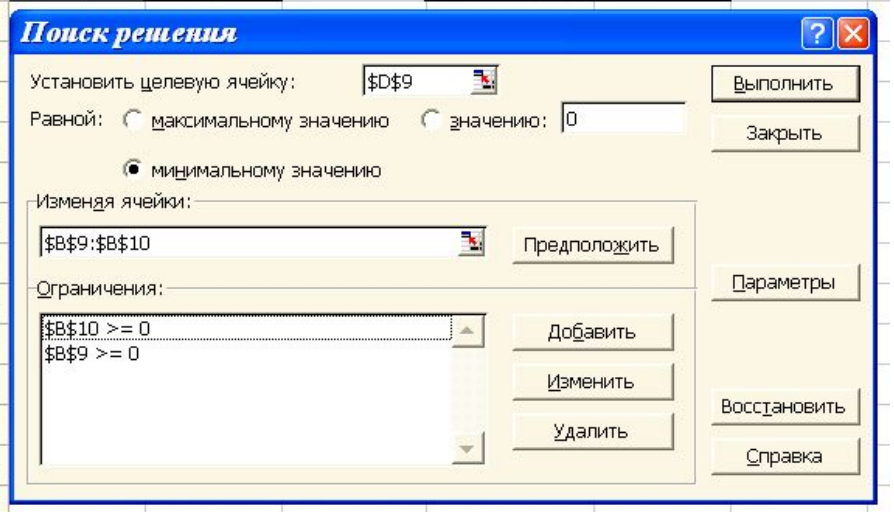
Степенная – $f(x) = \beta_0 x^{\beta_1}$.

Показательная – $f(x) = \beta_0 \beta_1^x$.

Логарифмическая – $f(x) = \beta_0 + \beta_1 \ln x$.

Нелинейная регрессия

	A	B	C	D	E	F
1	Исходные данные		$\hat{y}_i$	$(\hat{y}_i - y_i)^2$		
2	1	10	1	81,000		
3	2	13,4	1,231	148,081		
4	3	15,4	1,390	196,269		
5	4	16,5	1,516	224,529		
6	5	18,6	1,621	288,298		
7	6	19,1	1,712	302,351		
8						
9	b_0	1	$F(b_0, b_1)$	1240,528		
10	b_1	0,3		=СУММ(D2:D7)		
11						
12						
13						
14						
15						
16						
17						
18						
19						
20						
21						
22						
23						



$$\hat{y}_i = b_0 \cdot x_i^{b_1}, \quad i = 1,\ldots,6$$

в ячейке C2 программируется выражение =B9*A2^B10

Нелинейная регрессия

- В случае нелинейной зависимости между исследуемыми факторами, степень их взаимосвязи характеризуется индексом корреляции:

$$I_{xy} = \sqrt{1 - \frac{Q_e}{Q}}$$

где $Q_e = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$, $Q = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$, $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$, $\hat{y}_i$ - значение зависимой переменной Y, вычисленное по уравнению нелинейной регрессии при $x = x_i$. Очевидно, что величина этого показателя удовлетворяет неравенству: $0 \le I_{xy} \le 1$, причем $I_{xy} = 1$, когда все значения y_i “лежат” на линии регрессии.

Регрессия

На этапе регрессионного анализа решаются следующие задачи:

1. Выбор общего вида уравнения регрессии и определение параметров регрессии.

2. Определение степени взаимосвязи результативного признака и факторов, проверка общего качества уравнения регрессии.

3. Проверка статистической значимости каждого коэффициента уравнения регрессии и определение их доверительных интервалов.

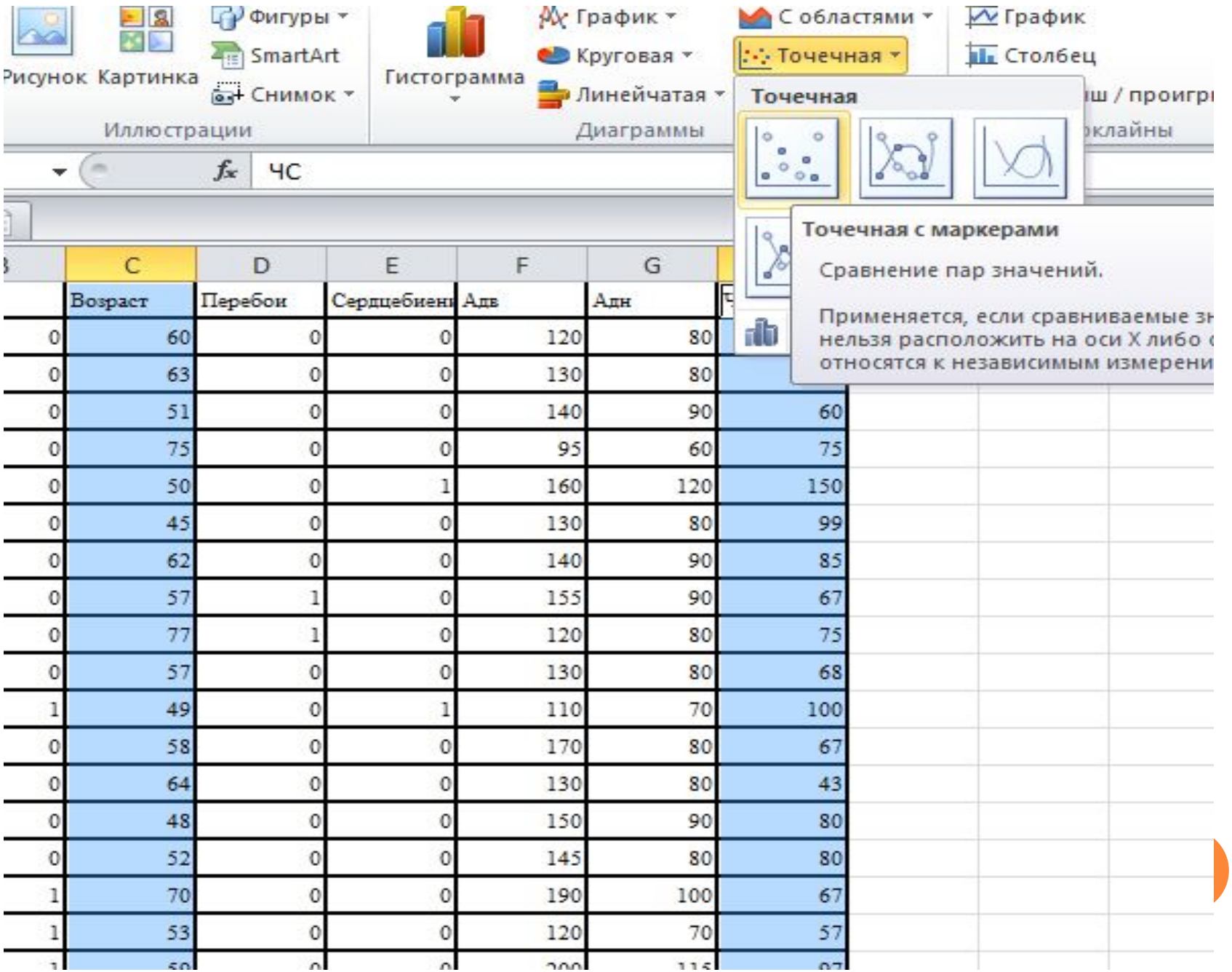
Уравнение простой линейной регрессии имеет вид: $y = b_0 + b_1 x$, множественная линейная регрессия описывается следующим уравнением: $y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_m x_m$.

Регрессия

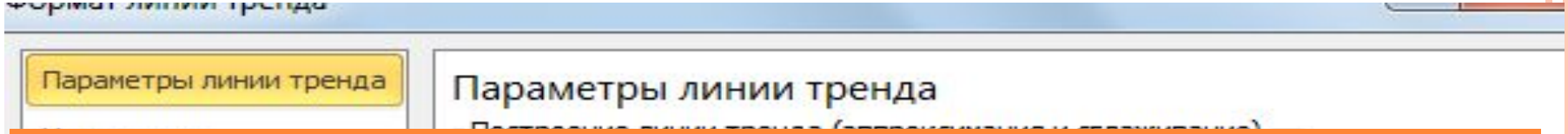
J2 =КОРРЕЛ(C2:C41;H2:H41)

Книга2 *

A	B	C	D	E	F	G	H	I	J
N	Пол	Возраст	Перебои	Сердцебиени	Адв	Адн	ЧС		
1	0	60	0	0	120	80	57		-0,17428918
2	0	63	0	0	130	80	60		
3	0	51	0	0	140	90	60		
4	0	75	0	0	95	60	75		
5	0	50	0	1	160	120	150		
6	0	45	0	0	130	80	99		
7	0	62	0	0	140	90	85		
8	0	57	1	0	155	90	67		
9	0	77	1	0	120	80	75		
10	0	57	0	0	130	80	68		
11	1	49	0	1	110	70	100		
12	0	58	0	0	170	80	67		
13	0	64	0	0	130	80	43		
14	0	48	0	0	150	90	80		
15	0	52	0	0	145	80	80		
16	1	70	0	0	190	100	67		



	C	D	E	F	G	
	Возраст	Перебои	Сердцебиени	Адв	Адн	
0	60	0	0	120	80	
0	63	0	0	130	80	
0	51	0	0	140	90	60
0	75	0	0	95	60	75
0	50	0	1	160	120	150
0	45	0	0	130	80	99
0	62	0	0	140	90	85
0	57	1	0	155	90	67
0	77	1	0	120	80	75
0	57	0	0	130	80	68
1	49	0	1	110	70	100
0	58	0	0	170	80	67
0	64	0	0	130	80	43
0	48	0	0	150	90	80
0	52	0	0	145	80	80
1	70	0	0	190	100	67
1	53	0	0	120	70	57



При проведении статистических исследований получаемые результаты часто представляются в виде упорядоченных последовательностей значений этих результатов, называемых элементами последовательности. Упорядочение заключается в том, что каждому элементу последовательности присваивается соответствующий номер. При этом полученные результаты записываются в порядке возрастания их номеров.

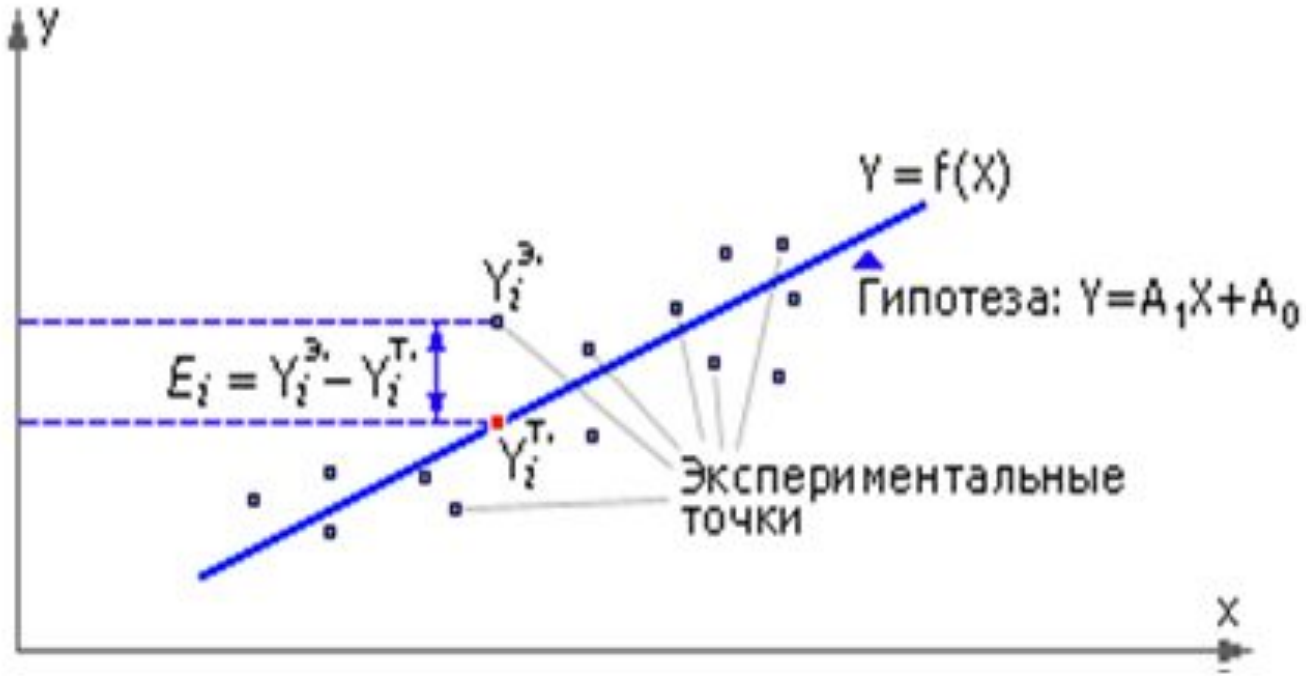
Временной ряд – это ряд последовательных значений, характеризующих изменение показателя во времени (показатели кардиограммы).

Трендом (trend – тенденция, направление) временного ряда называют изменяющийся, нециклический компонент, описывающий влияние долговременных факторов, эффект которых сказывается постепенно. К таким факторам относятся изменение демографических характеристик, рост рождаемости и др.

Сезонный компонент временного ряда описывает поведение, изменяющееся регулярно в течение заданного периода (года, месяца, недели, дня и т.п.). Состоит из почти повторяющихся циклов (пики сезонных заболеваний).

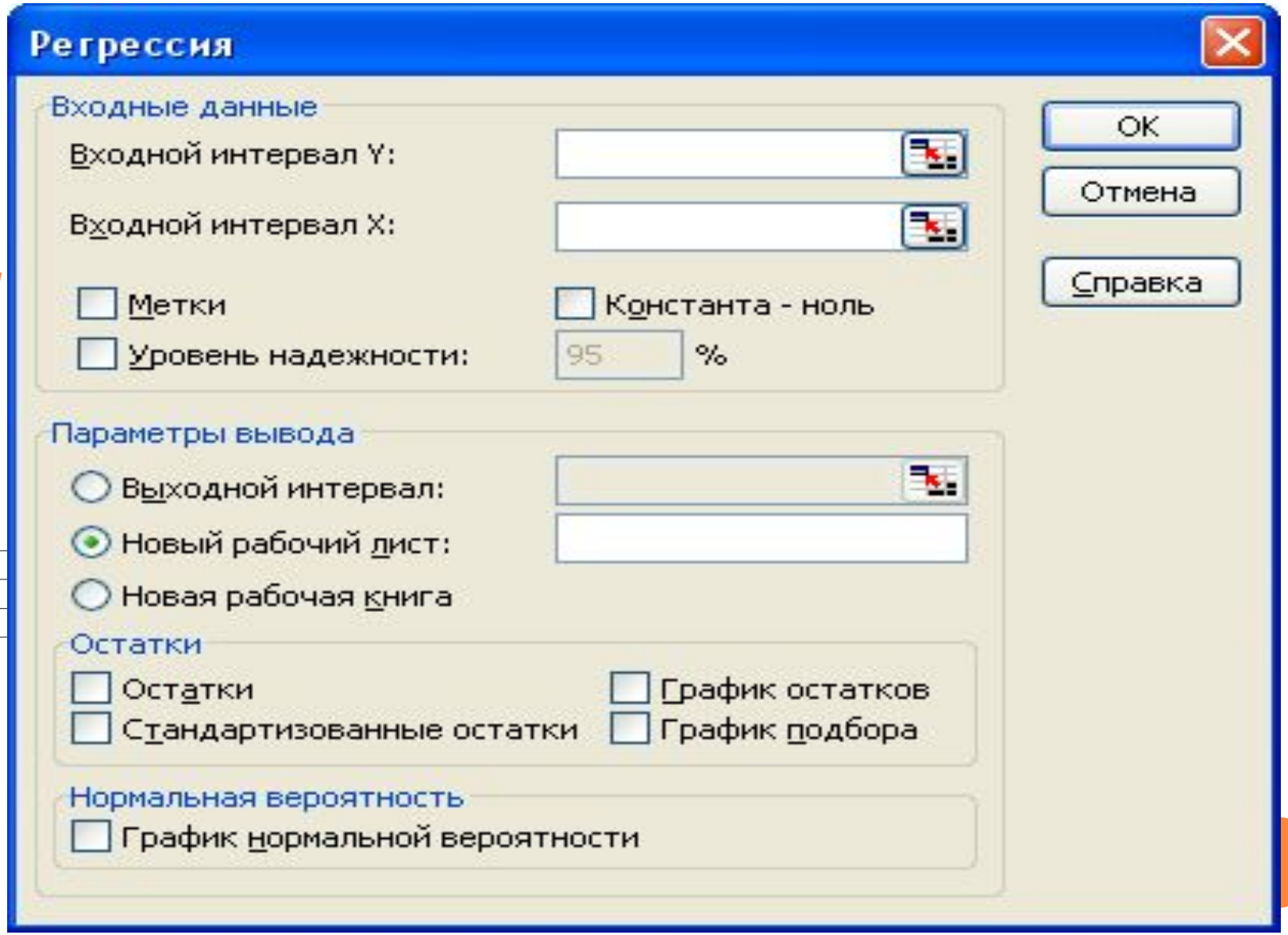
Циклический компонент описывает длительные периоды относительного подъема и спада и состоит из циклов, меняющихся по амплитуде и протяженности.

РЕГРЕССИОННЫЙ АНАЛИЗ



$$Q = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 \to \min .$$

РЕГРЕССИОННЫЙ АНАЛИЗ



Параметры

1. *Входной интервал* Y – вводится диапазон ячеек (**один столбец**), содержащих исходные данные по **результирующему признаку**.

2. *Входной интервал* X – вводится диапазон ячеек (**число столбцов равно количеству признаков**), содержащих исходные данные **факторного признака**.

3. *Метки* – флажок ставится, если первая строка содержит заголовок, в противном случае будут созданы стандартные заголовки автоматически.

4. *Уровень надежности* – флажок устанавливается, если требуется ввести значение уровня отличное от 95%. При выключенном флажке уровень надежности принимается равным 95%.

5. *Константа-ноль* - флажок устанавливается в том случае, когда требуется, чтобы линия регрессии прошла через начало координат, т.е. $b=0$

6. *Параметры вывода* – указывается место, где будут указаны таблицы результатов анализа.

7. *Остатки* – при необходимости вывода столбцов остатков и графиков остатков и подбора необходимо включить соответствующие флажки.

8. *Нормальная вероятность* – флажок устанавливается, если не требуется вывести график зависимости наблюдаемых значений от автоматически формируемых интервалов персентилей.

РЕГРЕССИОННЫЙ АНАЛИЗ

Регрессионная статистика	
Множественный R	0,984535285
R-квадрат	0,969309728
Нормированный R-квадрат	0,959079637
Стандартная ошибка	1,724792855
Наблюдения	5

Множественный R – коэффициент корреляции.
R-квадрат – коэффициент детерминации.
Нормированный R-квадрат – нормированное значение коэффициента корреляции.
Стандартная ошибка - стандартное отклонение для остатков.
Наблюдения - количество исходных наблюдений.

КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ

- Одной из *наиболее эффективных оценок адекватности уравнения регрессии* (мерой качества «подгонки» регрессионной модели к «наблюденным» значениям y_i) является *коэффициент детерминации* R^2, определяемый по формуле:

$$R^2 = \frac{Q_r}{Q} = 1 - \frac{Q_e}{Q}$$

объясненная (или факторная) сумма квадратов (в переводной литературе – RSS)

$$Q_r = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2,$$

$$Q_e = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2,$$

являющаяся мерой разброса (разброса точек относительно линии регрессии), не «объясненного» построенным уравнением регрессии.

КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ

$$0 \le R^2 \le 1$$

- Чем ближе R^2 к 1, тем лучше регрессия аппроксимирует эмпирические данные. Если $R^2 = 1$, то эмпирические точки (x_i, y_i) лежат на линии регрессии ($Q_e = 0$), и между X и Y существует линейная функциональная зависимость. Если $R^2 = 0$ ($Q_e = Q$), то вариации Y полностью обусловлены воздействием неучтенных в уравнении регрессии переменных, и линия регрессии параллельна оси абсцисс.
- **Внимание!** Коэффициент R^2 имеет смысл рассматривать, если в уравнении регрессии *присутствует свободный член* (в случае парной линейной регрессии – коэффициент b_0).
- В случае парной линейной регрессии имеет место важное тождество

$$R^2 = r_{XY}^2$$

РЕГРЕССИОННЫЙ АНАЛИЗ

Дисперсионный анализ – анализ изменчивости результативного признака под влиянием каких-либо контролируемых переменных факторов.

Дисперсионный анализ — статистический метод, применяемый для выявления влияния отдельных факторов (количественных, порядковых или качественных) на изучаемый признак и оценку степени этого влияния.

df – число степеней свободы. Для строки *Регрессия* это количество факторных признаков, для строки *Остаток* – число наблюдений минус количество переменных в уравнении регрессии, для строки *Итого* – сумма степеней свободы для строк *Регрессия* и *Остаток*.

SS – сумма квадратов отклонений. Для строки *Регрессия* это значение определяется как сумма квадратов отклонений теоретических данных от среднего, для строки *Остаток* это сумма квадратов отклонений эмпирических данных от теоретических, для строки *Итого* это сумма квадратов отклонений эмпирических данных от среднего.

MS – дисперсии. Для строки *Регрессия* это факторная дисперсия, для строки *Остаток* это остаточная дисперсия.

F – расчетное значение *F*-критерия Фишера, определяемое как отношение факторной дисперсии к остаточной.

Значимость F – значение уровня значимости, соответствующее вычисленному значению *F*.

Дисперсионный анализ

- Если изучается действие количественного фактора, то предварительно производится его разбивка на градации. Для каждой градации подсчитывается среднее значение изучаемого признака, затем дисперсия среднего по градациям фактора относительно общего среднего и, наконец, общая дисперсия изучаемого показателя (независимо от значения фактора).
- В теории дисперсионного анализа показано, что общая дисперсия D равна дисперсии средних по градациям фактора D_F (доля дисперсии за счет действия исследуемого фактора — объясненная дисперсия) плюс остаточная дисперсия за счет действия случайных факторов (D_S):
- $D = D_F + D_S$.
- Чем больше эта величина, тем сильнее влияние фактора на изучаемый признак. Для количественной оценки степени влияния вычисляют показатель F по формуле:

$$F = \frac{D_F / (L-1)}{D_S / (N-L)}$$

- где L — число градаций фактора, N — объем генеральной совокупности.
- Показатель влияния F затем сравнивается со стандартным значением F_{st} в таблице Фишера (для выбранного уровня значимости при соответствующем числе степеней свободы). Если $F > F_{st}$ то факт влияния считается достоверно доказанным.

Дисперсионный анализ в системах имитационного моделирования

- Статистический метод анализа результатов наблюдений, зависящих от различных, одновременно действующих факторов, выбор наиболее важных факторов и оценка их влияния.
- С помощью него определяются количественные отклонения наблюдений от средних значений.
- Если какой-либо фактор не оказывает влияния на отклик, то он является **незначимым**.
- ***Главным эффектом фактора j*** называется средняя величина изменения в отклике, обусловленная переходом фактора j с уровня « – » на уровень «+», в то время как остальные факторы остаются без изменений.

Дисперсионный анализ

- ***Эффектом взаимодействия*** можно назвать комбинированное влияние на отклик двух или более факторов, проявляющееся помимо индивидуального влияния всех этих факторов по отдельности.

- ***Эффект взаимодействия*** определяется как половина разности между средним эффектом фактора j_1, когда фактор j_2 находится на уровне «+» (а все остальные факторы, кроме j_1 и j_2 остаются без изменений) и средним эффектом фактора j_1, когда фактор j_2 находится на уровне «-».

ГЛАВНЫЙ ЭФФЕКТ ФАКТОРА J

Точка плана	Фактор 1	Фактор 2	Фактор 3	Отклик
1	-	-	-	y1
2	+	-	-	y2
3	-	+	-	y3
4	+	+	-	y4
5	-	-	+	y5
6	+	-	+	y6
7	-	+	+	y7
8	+	+	+	y8

первого:

$e_1 = ((y_2-y_1)+(y_4-y_3)+(y_6-y_5)+(y_8-y_7)) / 4$

второго:

$e_2 = ((y_3-y_1)+(y_4-y_2)+(y_7-y_5)+(y_8-y_6))/4$

третьего:

$e_3 = ((y_5-y_1)+(y_6-y_2)+(y_7-y_3)+(y_8-y_4))/4.$

ЭФФЕКТЫ ВЗАИМОДЕЙСТВИЯ

$$e_{12} = \frac{1}{2}\cdot\left(\frac{y_4 - y_3 + y_8 - y_7}{2} - \frac{y_2 - y_1 + y_6 - y_5}{2}\right)$$

$$e_{13} = \frac{1}{2}\cdot\left(\frac{y_6 - y_5 + y_8 - y_7}{2} - \frac{y_2 - y_1 + y_4 - y_3}{2}\right)$$

$$e_{23} = \frac{1}{2}\cdot\left(\frac{y_7 - y_5 + y_8 - y_6}{2} - \frac{y_3 - y_1 + y_4 - y_2}{2}\right)$$

Факторный анализ

- **Факторный анализ** — совокупность методов исследования многомерных признаков за счет снижения их размерности (путем введения так называемых общих факторов, которые непосредственно наблюдаться не могут). В медицине методы факторного анализа применяются для решения двух взаимосвязанных задач: группировки исходной системы признаков на основе их корреляционных связей и сжатия информации за счет построения системы обобщенных индикаторов.
- В факторной модели каждый исходный признак представляется в виде комбинации новых показателей (общих факторов), число которых, как правило, устанавливается меньше числа исходных. Такой метод описания удобен, например, для получения обобщенных индексов, характеризующих состояние системы здравоохранения различных регионов или однородных учреждений (исходные показатели — заболеваемость, смертность, количество профосмотров — заменяются набором обобщенных показателей, определяющих ресурсное обеспечение, качество врачебного обслуживания и т.п.).
- Недостатком факторного анализа является трудность содержательной интерпретации общих факторов.

РЕГРЕССИОННЫЙ АНАЛИЗ

Дисперсионный анализ					
	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1	281,8752688	281,8752688	94,75084337	0,002303227
Остаток	3	8,924731183	2,974910394		
Итого	4	290,8			

Столбец df - число степеней свободы. Для строки *Регрессия* показатель равен числу независимых переменных $k_r = k = m\text{-}1$;
для строки *Остаток* - равен $k_o = n - (k_r + 1) = n\text{-}m$;
для строки *Итого* – равен $k_r + k_o$
Столбец F – значение *Fc* , равное *F* критерию Фишера
Столбец значимость F - значение уровня значимости, соответствующее вычисленной величине *F* критерия и равное вероятности $P(F(k_r, k_o) \geq F_c)$, где $F(k_r, k_o)$ - случайная величина, подчиняющаяся распределению Фишера с *kr* ,*ke* степенями свободы. Эту вероятность можно также определить с помощью функции = FРАСП($Fc; k_r; k_e$).
Если вероятность меньше уровня значимости α *(обычно* 0.05 *), то построенная регрессия является значимой.*

РЕГРЕССИОННЫЙ АНАЛИЗ

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересечение	-1,064516129	1,710826692	-0,622223241	0,577881718
срок службы	2,752688172	0,282790929	9,734004488	0,002303227

Коэффициенты – значения коэффициентов модели.

Стандартная ошибка – стандартные ошибки коэффициентов.

t-статистика – расчетные значения *t*-критерия, вычисляемого как отношение значений коэффициентов к соответствующим стандартным ошибкам.

P-Значение – значения уровней значимости, соответствующие вычисленным значениям t_p.

Помимо этого указываются нижние и верхние границы доверительных интервалов для коэффициентов регрессии - *Нижние 95%, Верхние 95%*

Для проверки значимости коэффициентов сформулируем *статистические гипотезы*:

H_0: коэффициент b_0 не значим

H_1: коэффициент b_0 значим

и примем уровень значимости (вероятность ошибки первого рода) равным = 0.05.

Если вероятность P-значение меньше уровня значимости , то принимается гипотеза о значимости соответствующего коэффициента регрессии.

РЕГРЕССИОННЫЙ АНАЛИЗ

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересечение	-1,064516129	1,710826692	-0,622223241	0,577881718
срок службы	2,752688172	0,282790929	9,734004488	0,002303227

На основе данных из полученных таблиц можно сделать следующие выводы:

1. Уравнение регрессии имеет вид: **. *Y=–1,06*+2,75 x**
2. Значение коэффициента детерминации, равного 0,97 показывает, что срок службы существенно влияют на затраты на ТО, что подтверждает правильность включения его в построенную модель.
3. Рассчитанный уровень значимости ***Значимость F* = 0,002 меньший 0,05** подтверждает **значимость величины коэффициента детерминации**.
4. *P-Значение* для срока службы, равное **0,002 и меньшее 0,05** подтверждает значимость коэффициента b_1
5. *P-Значение* для коэффициента **превышает 0,05**, это означает, что данный коэффициент для модели не является значимым и его можно опустить, т.е. график модели будет проходить через точку начала координат b_0

Функции Excel для регрессионного анализа

Статистические функции Excel, полезные при построении парной линейной регрессии.

- ***Функция ОТРЕЗОК.*** Вычисляет коэффициент $b0$ и обращение имеет вид

ОТРЕЗОК(*диапазон_значений_ y ; диапазон_значений_ x*).

- ***Функция НАКЛОН.*** Вычисляет коэффициент $b1$ и обращение имеет вид

НАКЛОН(*диапазон_значений_ y ; диапазон_значений_ x*).

Функции Excel для регрессионного анализа

- ***Функция ПРЕДСКАЗ.*** Вычисляет значение линейной парной регрессии при заданном значении независимой переменной (обозначена через *z*) и обращение имеет вид

ПРЕДСКАЗ(*z;диапазон_значений_y;диапазон_значений_ x*).

- Функция ***ТЕНДЕНЦИЯ*** возвращает значения в соответствии с линейным трендом. Аппроксимирует прямой линией (по методу наименьших квадратов) массивы *известные_значения_y* и *известные_значения_x*. Возвращает значения y, в соответствии с этой прямой для заданного массива *новые_значения_x*.

ТЕНДЕНЦИЯ(y; x; n_x; конст):

- y - *известные_значения_y* – множество значений y, для которых уже известна линейная зависимость;
- x - *известные_значения_x* - множество значений x, для которых уже известна линейная зависимость;
- n_x - *новые_значения_x* – новые значения x, для которых функция возвращает соответствующие значения y.
- *конст* – логическое значение, если оно равно 0, то свободный член равен нулю, в противном случае свободный член вычисляется обычным образом.

Регрессионный анализ данных

Функция ***РОСТ*** рассчитывает прогнозируемый экспоненциальный рост на основании имеющихся данных:

РОСТ(y; x; n_x; конст):

- y - *известные_значения_y* – множество значений y, для которых уже известна экспоненциальная зависимость;
- x - *известные_значения_x* - множество значений x, для которых уже известна экспоненциальная зависимость;
- n_x - *новые_значения_x* – новые значения x, для которых функция возвращает соответствующие значения y.
- *конст* – логическое значение, если оно равно 0 или отсутствует, то константа равна единице, в противном случае вычисляется обычным образом.

РЕГРЕССИОННЫЙ АНАЛИЗ

- $$R_{x/y}^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$, где $\hat{y}_i$ - расчетное, y_i - измеренное $\bar{y}$ -среднее

  детерминации.
- Проведя расчеты, основанные на одних и тех же исходных данных, для нескольких типов функций, мы можем из них выбрать такую, которая дает наибольшее значение R^2
- Чем больше R^2, т. е. чем больше числитель, тем больше изменение факторного признака объясняет изменение результативного признака и тем, следовательно, лучше уравнение регрессии, лучше выбор функции.

СПАСИБО ЗА ВНИМАНИЕ.