РЕГРЕССИОННЫЙ АНАЛИЗ

В практике статистического исследования весьма часто возникает необходимость определить не только корреляционное соотношение между изучаемыми характеристиками, но и установить определенную обусловленность между ними, представив выявленную связь в строгой аналитической форме. В этом случае результат исследования – экспериментальная зависимость воздействия какого-либо фактора (скажем, производительности труда, уровня образования, практического стажа работы и т.д.) на изменение изучаемого параметра (например, величины прибыли фирмы) − может быть не только представлен в виде графика (что весьма наглядно), но и описан математически с использованием аппроксимирующего выражения (эмпирической формулы). Исследование такой ситуации и является задачей регрессионного анализа, который дает предсказание (прогнозирование) одной переменной на основании другой.

Для выполнения такого прогноза требуется определить коэффициент корреляции Пирсона, с использованием которого вычисляют коэффициент регрессии ($\boldsymbol{R_{y/x}}$). Он участвует в создании регрессионной функции вида $\boldsymbol{y=ax+b}$, которая применяется для прогнозирования требуемых параметров.

Коэффициент регрессии вычисляется по формуле:

$$\boldsymbol{R_{y/x} = r_{x/y}\ \frac{\sigma_y}{\sigma_x}}\ ,$$

где: $\boldsymbol{R_{y/x}}$ – коэффициент регрессии;
$\boldsymbol{r_{x/y}}$ – коэффициент корреляции Пирсона;
$\boldsymbol{\sigma_x}$ – среднее квадратическое отклонение признака $\boldsymbol{x}$;
$\boldsymbol{\sigma_y}$ – среднее квадратическое отклонение признака $\boldsymbol{y}$.

Среднее квадратическое отклонение (сигма) вычисляется по формуле:

$$\sigma = \sqrt{\frac{\sum d^2}{n}}\ ,$$

а в программе Excel функцией = СТАНДОТКЛОН(Диапазон ячеек).

Значение коэффициента регрессии ($\boldsymbol{R_{y/x}}$) в программе Excel может быть вычислено функцией =НАКЛОН(Диапазон_y; Диапазон_x).

Формула определения значения зависимого признака:

$$\boldsymbol{y = M_y + R_{y/x}\,(x - M_x)}\ ,$$

где: $\boldsymbol{y}$ – зависимая переменная;
$\boldsymbol{M_y}$ – средняя признака $\boldsymbol{y}$;
$\boldsymbol{R_{y/x}}$ - коэффициент регрессии;
$\boldsymbol{x}$ - значение измеренного признака;
$\boldsymbol{M_x}$ – средняя арифметическая признака $\boldsymbol{x}$.

В программе Excel значение зависимой переменной (***y***) при заданном значении ***x*** может быть вычислено функцией =ПРЕДСКАЗ(***x*** ; Диапазон_y; Диапазон_x).

После получения прогнозируемого значения (***y***) выполняется определение его доверительного интервала с целью экстраполяции данных на генеральную совокупность с уровнем значимости $p<0,05$. Для этого вычисляется сигма регрессии $\boldsymbol{\sigma}_{Ry/x}$, которая показывает меру вариабельности зависимого признака, вычисленного по уравнению регрессии, в генеральной совокупности.

Она определяется по формуле: $\boldsymbol{\sigma}_{Ry/x} = \boldsymbol{\sigma}_y \sqrt{1 - \mathbf{r}_{xy}^2}$. Вычисление значения $\boldsymbol{\sigma}_y$ может производиться функцией = СТАНДОТКЛОН(Диапазон_y).

<u>Пример</u> прогнозирования значения одного признака по известному значению другого с помощью уравнения регрессии.

<u>Условие задачи:</u> на основе данных, характеризующих уровень запыленности рабочих мест, необходимо выполнить прогноз уровня пыли при температуре воздуха $23C^0$.

Результаты измерений запыленности в помещении с учетом температуры окружающей среды

Измерение на рабочем месте	Температура воздуха C^o	Запыленность мг/м^3
1. Слесарь	20	0,2
2. Электрик	21	0,25
3. Сварщик	21	0,24
4. ...	19	0,08
5. ...	19	0,08
6. ...	19	0,07
7. ...	22	0,3
8. ...	22	0,28
9. ...	25	0,33
10....	24	0,31
11....	21	0,26
12....	21	0,27

<u>Задание:</u> построить уравнение регрессии для зависимости между температурой окружающей среды и уровнем запыленности помещения, создать регрессионную функцию и вычислить значение уровня пыли при температуре воздуха $23C^o$. Определить сигму регрессии и доверительный интервал для прогнозируемого значения уровня пыли.

Решение: запустите программу Excel. Переименуйте Лист 1, обозначив названием «Регрессия». На этом листе введите данные и решение задачи, как показано ниже.

а) первоначально требуется выполнить вычисление коэффициента корреляции Пирсона с помощью таблицы отклонений (см рис.1)

Вычисление коэффициента корреляции Пирсона							
Варианта	Температура воздуха (x)	Запыленность мг/м3 (y)	d_x	d_y	d_x*d_y	d_x^2	d_y^2
1	19	0,07	2,2	0,153	0,330	4,7	0,0233
2	19	0,08	2,2	0,143	0,309	4,7	0,0203
3	19	0,08	2,2	0,143	0,309	4,7	0,0203
4	20	0,2	1,2	0,023	0,026	1,4	0,0005
5	21	0,24	0,2	-0,018	-0,003	0,0	0,0003
6	21	0,25	0,2	-0,028	-0,005	0,0	0,0008
7	21	0,26	0,2	-0,038	-0,006	0,0	0,0014
8	21	0,27	0,2	-0,048	-0,008	0,0	0,0023
9	22	0,3	-0,8	-0,078	0,065	0,7	0,0060
10	22	0,28	-0,8	-0,058	0,048	0,7	0,0033
11	24	0,31	-2,8	-0,088	0,248	8,0	0,0077
12	25	0,33	-3,8	-0,108	0,412	14,7	0,0116
Средняя (M) =	21,2	0,223		Сумма (Σ) =	1,725	39,7	0,0976
n=	12						
$r_{xy} = \frac{\sum d_x d_y}{\sqrt{\sum d_x^2 \sum d_y^2}}$		r_{xy}=	0,88				

Рис.1

$$r_{xy} = \frac{\sum d_x d_y}{\sqrt{\sum d_x^2 \sum d_y^2}} = \frac{1{,}725}{\sqrt{39{,}7 \times 0{,}0976}} = \underline{\mathbf{0{,}88}}\ .$$

б) вычисление коэффициента регрессии ($R_{y/x}$) (см рис. 2):

Сигма $\sigma = \sqrt{\frac{\sum d^2}{n}}$ $\quad \sigma_x =$ КОРЕНЬ(39,7/12) =СТАНДОТКЛОН(x_1:x_n) = 1,90

$\sigma_y =$ КОРЕНЬ(0,0976/12) =СТАНДОТКЛОН(y_1:y_n) = 0,09

$R_{y/x} = r_{x/y} \frac{\sigma_y}{\sigma_x}$ $\quad R_{y/x}=$ 0,88*0,09/1,9 = ***0,04***

Вычисление коэффициента регрессии:		
$\sigma = \sqrt{\frac{\sum d^2}{n-1}}$	δ_x =	1,90
	δ_y =	0,09
$R_{y/x} = r_{x/y} \frac{\sigma_y}{\sigma_x}$	$R_{x/y}$ =	0,04

Рис. 2

в) вычисление величины зависимого признака (***y***) при температуре 23С^0 (см. рис. 3):

$y = M_y + R_{y/x}(x - M_x)$ При x= 23С^0

y = 0,223+0,04(23-21,2) =ПРЕДСКАЗ(x ; y_1:y_n ; x_1:x_n) = ***0,30*** мг/м^3

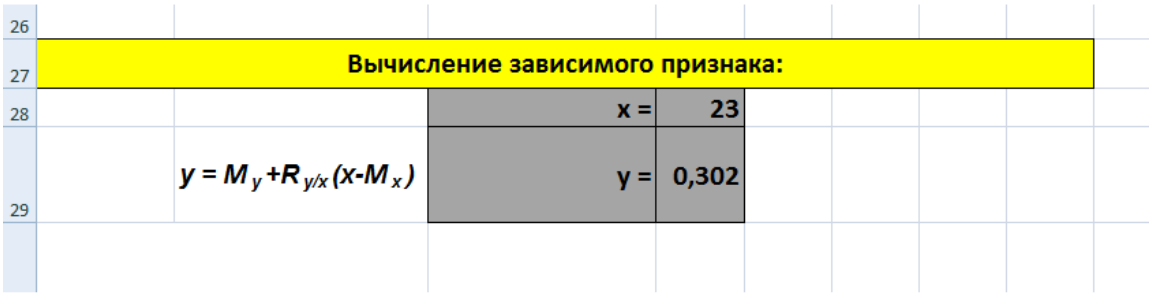

Рис. 3

г) вычисление доверительных границ колебаний зависимого признака в генеральной совокупности (см. рис. 4):

$\sigma_{Ry/x} = \sigma_y \sqrt{1 - r_{xy}^2}$ $\sigma_{Ry/x}$ = 1,9*КОРЕНЬ(1-0,88^2) = ***0,045*** .

Доверительные границы: *от* ***0,21*** *до* ***0,39*** При p<0/05

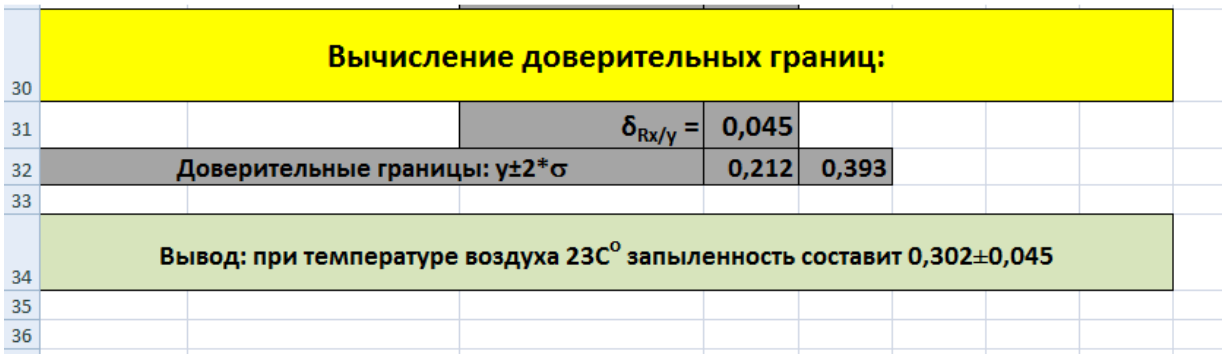

Рис. 4

Вывод: при температуре воздуха 23С° запыленность составит от 0,21 до 0,39мг/м3.

Практические задания.

☝ Задания выполняются по вариантам. Прежде чем приступить к выполнению задания, уточните свой вариант у преподавателя.

Задание 1. Задача прогнозирования значения одного признака по известному значению другого с помощью уравнения регрессии.

Ход выполнения задания

1. Запустите программу Excel

2. Переименуйте Лист 1, обозначив название «Регрессия».
3. Вычислите коэффициент корреляции Пирсона.
4. Вычислите коэффициент регрессии и постройте уравнение регрессии.
5. Вычислите величину зависимого признака.
6. Вычислите доверительные границы колебаний зависимого признака в генеральной совокупности. Сделайте вывод.
7. Сохраните файл.

Вариант 1

Представлены стандартизированные показатели (на 100000 населения) заболеваемости раком легкого с 1990 по 2005 г и выбросы загрязняющих веществ (ЗВ) в атмосферу (тыс.т.). Постройте уравнение регрессии для зависимости заболеваемостью раком и выбросами ЗВ. Определите заболеваемость раком легкого если выброс ЗВ составит 1700 тыс. т., вычислите сигму регрессии и доверительный интервал для полученного значения заболеваемости.

№	*Выбросы ЗВ(тыс.т.)*	*Заболеваемость раком легкого*
1	2060,300	48
2	2081,500	53
3	1894,800	55
4	1805,217	48
5	1670,194	45
6	1679,813	48
7	1515,861	51
8	1221,827	46
9	1416,746	44
10	1350,032	40
11	1283,300	40
12	1254,434	40
13	1191,775	39,1

Вариант 2

Выполнены измерения уровня заболеваемости органов дыхания на вредных производствах от содержания в воздухе двуокиси углерода. Постройте уравнение регрессии для зависимости между уровнем заболеваемости и содержание в воздухе двуокиси углерода. Определите значение уровня заболеваемости если содержание в воздухе двуокиси углерода составит 1,4. Вычислите сигму регрессии и доверительный интервал для полученного значения заболеваемости.

Уровень заболеваемости	Содержание в воздухе двуокиси углерода
1160	0,78
1155	0,88
1158	1,1
1157	1
1160	0,9
1161	0,9
1157	0,88
1159	0,75
1256	1,2
1260	1,2
1040	0,6
1039	0,6

Вариант 3

Изучена еженедельная заболеваемость острыми респираторными инфекциями на территории Н. в зимний период (с декабря по февраль) в зависимости от средней еженедельной температуры воздуха. Постройте уравнение регрессии для зависимости между количеством ОРЗ и температурой воздуха. Определите количество заболеваний если средняя температура составит -15^0C. Вычислите сигму регрессии и доверительный интервал для полученного значения заболеваемости.

Неделя	***Количество ОРЗ***	***Температура воздуха -t^0C***
1	30	20
2	31	21
3	33	22
4	34	23
5	34	21
6	36	25
7	38	25
8	39	29
9	38	28
10	36	23
11	28	20
12	34	22

Вариант 4

Под влиянием облучения рентгеновскими лучами наблюдалось следующее замедление размножения вируса мозаики Аукуба (в тыс.) в зависимости от длительности облучения (в мин.) Постройте уравнение регрессии для зависимости между временем облучения и размножением вируса. Определите количество вирусных частиц если время облучения составит 1 час 20 мин. Вычислите сигму регрессии и доверительный интервал для полученного значения количества вирусных частиц.

Количество вирусных частиц (в тыс.)	**Длительность облучения (мин.)**
271	0
226	3
209	7,5
190	10
158	15
129	21
105	30
69	45
32	53
19	60

Вариант 5

В результате исследования были получены следующие данные о весе щитовидной железы (в граммах) и соответствующей ей площади скеннографического изображения (в $см^2$). Постройте уравнение регрессии для зависимости между весом щитовидной железы и соответствующей ей площади скеннографического изображения. Определите вес щитовидный железы при площади скеннографического изображения, равной 90 $см^2$. Вычислите сигму регрессии и доверительный интервал для полученного значения веса щитовидной железы.

Вес щитовидной железы (в гр.)	**Площадь скеннографического изображения (в $см^2$)**
12	11
59	32
62	33

95	44
102	46
23	17
203	73
270	89
122	52
41	25

Вариант 6

В результате исследования были получены следующие данные о концентрации углекислоты (CO_2) в альвеолярном воздухе и ударном сердечном выбросе во время операции под наркозом. Постройте уравнение регрессии для зависимости между показателем CO_2 и соответствующим ему ударном сердечном выбросе. Определите ударный сердечный выброс при концентрации углекислоты, равной 6,5. Вычислите сигму регрессии и доверительный интервал для полученного значения ударного выброса.

Показатель CO_2	**Ударный выброс**
6,2	45
5,6	48
6,1	44
4,5	54
5,4	50
4,8	51
4,7	51
3,2	60
5,4	47
5,3	50
5,5	48
4,5	52

Вариант 7

Исследуется связь между поглощенной дозой облучения (в Гр) и долей аберрантных клеток костного мозга (в %) у подопытных животных. Постройте

уравнение регрессии для зависимости между дозой облучения и долей аберрантных клеток костного мозга. Определите значение необходимой дозы облучения если доля аберративных клеток костного мозга составит 100, вычислите сигму регрессии и доверительный интервал для полученного значения дозы облучения.

Номер наблюдения	***Доля аберративных клеток костного мозга, %***	***Доза облучения, Гр***
1	59	3,2
2	44	2,5
3	85	4,5
4	70	4,0
5	52	3,0
6	21	0,8
7	26	1,3
8	79	4,0
9	41	2,4
10	67	3,5
11	32	1,8
12	18	0,7
13	90	4,3
14	12	0,3

Вариант 8

В городе Н. было проведено изучение зависимости заболеваемости инфарктом миокарда по месяцам года в зависимости от среднемесячной температуры воздуха. Постройте уравнение регрессии для зависимости между среднемесячной температуры воздуха и уровнем заболеваемости инфарктом миокарда. Определите значение уровня заболеваемости инфарктом миокарда при температуре воздуха $+10\ C^0$. Вычислите сигму регрессии и доверительный интервал для полученного значения показателя заболеваемости.

Месяцы года	**Заболеваемость инфарктом миокарда по месяцам (на 10 000 жителей)**	**Среднемесячная температура воздуха**
январь	1,4	-8,5
февраль	1,23	-7,7
март	1,19	-5,8
апрель	1,13	-2,1
май	1	13
июнь	1,02	14,9

июль	0,91	18,8
август	1,02	15,6
сентябрь	1,06	9
октябрь	1,15	6
ноябрь	1,15	-1
декабрь	1,27	-7,7

Вариант 9

Изучалась зависимость между систолическим давлением мужчин в начальной стадии шока и возрастом. Постройте уравнение регрессии для зависимости между возрастом и систолическим давление. Определите значение систолического давления при возрасте пациента 60 лет. Вычислите сигму регрессии и доверительный интервал для полученного значения систолического давления.

Возраст	**Систолическое давление**
68	114
37	149
50	146
53	141
75	114
66	112
52	124
65	105
74	141
65	120
54	124

Вариант 10

В городе Н. было проведено изучение зависимости заболеваемости уролитиазом с 1998 по 2008 г. в зависимости от жесткости питьевой воды (моль/л). Постройте уравнение регрессии для зависимости между заболеваемостью

населения и жесткостью питьевой воды. Определите значение уровня заболеваемости уролитиазом при жесткости воды 5,0 моль/л. Вычислите сигму регрессии и доверительный интервал для полученного значения показателя заболеваемости.

Годы	**Жесткость воды, моль/л**	**Заболеваемость уролитиазом (на 1000 чел.)**
1998	4,8	0,6
2000	4,6	0,6
2001	5	0,67
2002	5,3	1
2003	5,8	1,46
2004	5,6	1,1
2005	5,9	1,6
2006	5,7	1,45
2007	5,4	1
2008	5,7	1,2

Вариант 11

У 22 женщин, преподавателей среднеобразовательной школы проведено изучение зависимости систолического артериального давления (САД) (мм.рт.мт.) от стажа работы в школе. Постройте уравнение регрессии для зависимости между САД и стажем работы. Определите значение артериального давления, если стаж работы составит 25 лет. Вычислите сигму регрессии и доверительный интервал для полученного значения показателя заболеваемости.

Систолическое артериальное давление, мм.рт.ст.	**Стаж преподавательской работы (годы)**
110	2
100	4
110	7
170	35

110	4
110	5
90	7
110	9
90	8
110	9
160	22
110	30
100	3
105	11
100	8
130	8
125	8
120	13
110	9
160	30
145	32
150	41

Вариант 12

В ходе клинических испытаний проведено изучение зависимости частоты сердечных сокращений (ЧСС) (уд/мин) от температуры тела (C^0). Постройте уравнение регрессии для зависимости между ЧСС и температурой. Определите значение ЧСС при температуре тела равной 41 C^0. Вычислите сигму регрессии и доверительный интервал для полученного значения показателя ЧСС.

Температура тела, C^0	**ЧСС, уд/мин**
36	90
36,5	95
39	120

38	110
36	85
40	140
37,5	100
38	110
36,6	95
38	110

КОНТРОЛЬНЫЕ ВОПРОСЫ:

1. Каково назначение регрессионного анализа?
2. В чем заключается задача построения регрессионной зависимости?
3. Условия и область применения регрессионного анализа.
4. Какой вид имеет линейное уравнение регрессии?
5. Методика расчета уравнения регрессии и сигмы регрессии.
6. Как построить линейную регрессионную модель в Excel?

Список литературы:

1) Зайцев В. М. Прикладная медицинская статистика: учеб. пособие для студентов мед. вузов / Зайцев В. М., Лифляндский В. Г., Маринкин В. И. . - СПб. : Фолиант , 2003 . - 430 с.

2) Ситуационные задачи по медицинской статистике с примерами решений в программе Microsoft Excel: учеб.-метод. пособие к практ. занятиям по дисциплине "Мед. информатика" для спец. : 060101 65 - Леч. дело, 060103 65 - Педиатрия, 060201 65 Стоматология, 060105 65 - Мед.-профил. дело / Голубев А. Н., Грибина Л. Н., Дьяченко Т. С. и др. ; ВолгГМУ Минздрава РФ . - Волгоград : Изд-во ВолгГМУ , 2014 . – 254с.

3) Сабанов В. И. Медицинская информатика и автоматизированные системы управления в здравоохранении: учеб.-метод. пособие к практ. занятиям / Сабанов В. И., Голубев А. Н., Комина Е. Р. ; Федерал. агентство по здравоохранению и соц. развитию; ВолГМУ . - Волгоград : Изд-во ВолГМУ , 2006 . - 144 с.

4) Голубев А.Н., Грибина Л.Н..Р., Бирюкова Л.Ф. и т.д. Тестовые задания по медицинской информатике и статистике. Учебное пособие / Под ред. проф. В.И. Сабанова. – Волгоград: Изд-во ВолгГМУ, 2014. – 426с.