МИНИСТЕРСТВО ЗДРАВООХРАНЕНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

Белорусская медицинская академия последипломного образования

Кафедра экономики и бухгалтерского учета в здравоохранении
с курсом медицинской информатики

Математическая статистика
Описательная статистика
Выполнение многомерного регрессионного анализа
(для аспирантов и соискателей ученой степени)

Учебно-методическое пособие

Минск БелМАПО
2016

УДК 579.22.23.25:61(075.9)
ББК 51.1(2)
М 35

Рекомендовано в качестве учебно-методического пособия
НМС Белорусской медицинской академии последипломного образования
протокол № 4 от 09.06. 2016г

Авторы:

Т.А. Радишевская ст. преподаватель,
Л.В. Шваб ст. преподаватель,
М.В.Щавелева доцент,
О.А.Кульпанович, доцент,
Ю.В. Мещеряков ст. преподаватель

Рецензенты:

Центр информатизации и инновационных разработок БГУИР
Радишевский В.А. – к.т.н., главный инженер проекта управления интегрированных систем ОДО «АВЕКТИС»

М 35 **Математическая** статистика, Описательная статистика, Выполнение многомерного регрессионного анализа: учеб.-метод. пособие /Т.А.Радишевская, Л.В. Шваб, [и др.].-–Минск : БелМАПО, 2016. - с.35

ISNB 978-985-458-044-3

Методические указания содержат краткое изложение необходимого теоретического материала и руководство по использованию регрессионного анализа в пакете statistica7.0..

Учебно-методическое пособие предназначено для аспирантов, ординаторов и соискателей ученой степени, обучающихся на кафедре экономики и бухгалтерского учета в здравоохранении с курсом медицинской информатикиБелМАПО.

УДК 579.22.23.25:61(075.9)
ББК 51.1(2)

ISBN 978-985-458-044-3

СОДЕРЖАНИЕ

ПРЕДИСЛОВИЕ

Методическое пособие по статистическому анализу исследований в физиологии предназначено для аспирантов и соискателей ученой степени, чтобы помочь начинающим исследователям освоить основные понятия математической статистики и более полно представить диапазон применения статистических методов.

В пособии рассматриваются основные статистические понятия, даны краткие описания методов обработки и анализа эмпирических данных: построение статистических оценок, параметрические и непараметрические методы проверки статистических гипотез, регрессионный анализ, некоторые вопросы планирования эксперимента.

Пособие начинается с обзора элементарных (основных) понятий математической статистики, а затем более подробно на конкретных примерах из различных областей медицины, включая лабораторные и другие виды исследований, показаны отдельные возможности и области использования программного пакета статистической диалоговой системы «*STATISTICA»*.

Представленный материал иллюстрирован рисунками. В приложении приведены таблицы основных статистических критериев.

1. ВВЕДЕНИЕ В СТАТИСТИКУ. ПОНЯТИЕ СТАТИСТИКИ

1.1. Предмет и метод статистической науки

Статистический показатель - это количественная оценка свойства изучаемого явления.

Оценочные показатели - это количественные характеристики изучаемых явлений на определенный момент времени (объемы, уровни).

Аналитические показатели - это показатели (относительные, средние величины, показатели вариации и динамики и др.), используемые для характеристики развития изучаемого явления (соотношение его отдельных частей, скорость развития во времени и т.д.)

Признак - характерное свойство изучаемого явления, отличающего его от других явлений, и выражается смысловыми понятиями и числовыми значениями.

Статистическая совокупность - это множество единиц изучаемого явления, объединенных в соответствии с задачей исследования.

Цель наблюдения - основной результат статистических исследований.

Объект статистического наблюдения - совокупность единиц изучаемого явления, о которых должны быть собраны статистические данные.

Единица наблюдения - это первичная ячейка, от которой должны быть получены необходимые статистические сведения.

Программа статистического наблюдения - перечень показателей, подлежащих изучению, регистрации по каждой единице наблюдения.

Статистическая методология - совокупность правил, приемов и методов статистического исследования. Статистическое исследование состоит из трех стадий:

1. Сбор первичной статистической информации (применяется **метод статистического наблюдения).**
2. Статистическая сводка и обработка первичной информации (собранная информация обрабатывается **методом статистических группировок** для выделения в изучаемой совокупности**).**
3. Анализ статистической информации. Обобщение и интерпретация (проводится **анализ статистической информации** на основе применения обобщающих статистических показателей: средних величин, вариации, скорости изменения явлений во времени, индексов и др).
4. **Статистическое наблюдение** - сбор первичных данных о изучаемом явлении. Осуществляется с помощью оценки и регистрации значений признаков единицы изучаемой совокупности в учетных документах (формы

отчетности предприятий, записи счетчиков в переписных листах ответов, и др.).

5. **Статистическая информация** (статистические данные) - первичный статистический материал, формирующийся в процессе статистического наблюдения, который подвергается систематизации, сводке, обработке, анализу и обобщению.

1.2.Принципы построения статистических группировок

Построение группировок начинают с выбора группировочного признака. Признаки различаются:

- по форме (атрибутивные и количественные);
- по степени и характеру колеблемости:
- по взаимосвязи (факторные признаки (х) воздействуют на другие, результативные (у) - зависят от других).

Количество групп для количественного признака задают по формуле_Стерджесcan = 1+3,322 lgN, где**n** -число групп, N - численность совокупности.

1.3.Статистические ряды распределения

После определения группировочного признака и границ групп строится ряд распределения.

Статистический ряд распределения - это упорядоченное распределение единиц совокупности на группы по группировочному признаку. Ряды распределения характеризуют структуру изучаемого явления, закономерности развития и т.д.

Ряды распределения образованные по атрибутивным признакам называются **атрибутивными** (распределение населения по полу), а по количественному - **вариационными** (распределение населения по возрасту).

Вариационные ряды, в зависимости от группировки по дискретному или непрерывному признаку, бывают **дискретными** или **интервальными.** Вариационные ряды состоят из 2-х элементов: **варианты и частоты.**

Варианта - числовое значение количественного признака в ряду распределения.

Частота - численности отдельных вариант или групп вариационного ряда (fi).

Сумма частот составляет **объем ряда** распределения V f=*n*. Частоты, выраженные в долях единицы или в процентах к объему ряда, называются**частостями.**

Полигоном графически изображаются дискретные ряды Интервальные ряды распределения графически отображаются **гистограммами. Гистограмма** - это график в виде вертикальных прямоугольников, высота которых соответствует их частотам, а ширина - величине интервала.

Кумулятивные ряды распределения строятся по накопленным частотам и показывают, сколько единиц совокупности, или какая их доля не превышает данное значение.

1.4.Нормальное распределение

Параметры:

- μ - коэффициент сдвига (вещественное число)
- σ > 0 - коэффициент масштаба (вещественный, строго положительный)
- Носитель
- Плотность вероятности
- Функция распределения
- Математическое ожидание
- Медиана
- Мода
- Дисперсия
- Коэффициент асимметрии
- Коэффициент эксцесса
- Информационная энтропия
- Производящая функция моментов
- Характеристическая функция

Нормальное распределение, также называемое гауссовским распределением или распределением Гаусса — распределение вероятностей, которое играет важнейшую роль во многих областях знаний. Исследуемая величина подчиняется нормальному распределению, когда она подвержена влиянию огромного числа случайных помех. Ясно, что такая ситуация крайне распространена, поэтому можно сказать, что из всех распределений в природе

чаще всего встречается именно нормальное распределение — отсюда и произошло одно из его названий.

Нормальное распределение зависит от двух параметров — смещения и масштаба, то есть является с математической точки зрения не одним распределением, а целым их семейством. Значения параметров соответствуют значениям среднего (математического ожидания) и разброса (стандартного отклонения).

Стандартным нормальным распределением называется нормальное распределение с математическим ожиданием 0 и стандартным отклонением 1.

Свойства

Если случайные величины X1 и X2 независимы и имеют нормальное распределение с математическими ожиданиями μ1 и μ2 и дисперсиями и соответственно, то X1 + X2 также имеет нормальное распределение с математическим ожиданием μ1 + μ2 и дисперсией .

1.5.Моделирование нормальных случайных величин

Простейшие, но неточные методы моделирования основываются на центральной предельной теореме. Именно, если сложить много независимых одинаково распределённых величин с конечной дисперсией, то сумма будет распределена примерно нормально. Например, если сложить 12 независимых базовых случайных величин, получится грубое приближение стандартного нормального распределения. Тем не менее, с увеличением слагаемых распределение суммы стремится к нормальному.

Статистическая проверка принадлежности к нормальному распределению

Поскольку нормальное распределение часто встречается на практике, то для него разработаны специальные статистические критерии проверки на «нормальность»:

1. Критерий Пирсона
2. Критерий Колмогорова-Смирнова
3. Критерий Шапиро-Вилка
4. График нормальности — не столько критерий, сколько графическая иллюстрация: точки специально построенного графика должны лежать почти на одной прямой.

1.6.Центральная предельная теорема

Нормальное распределение часто встречается в природе, нормально распределёнными являются следующие случайные величины:

- рост человека
- вес человека
- IQ

Такое широкое распространение закона связано с тем, что он является предельным законом, к которому приближаются многие другие (например, биномиальный).

Доказано, что сумма очень большого числа случайных величин, влияние каждой из которых близко к 0, имеет распределение, близкое к нормальному. Этот факт является содержанием центральной предельной теоремы.

1.7.Вычисление описательных статистик в системе "Statistica 7.0."

К числу описательных статистик относятся: среднее, выборочное среднее (mean), выборочная дисперсия (variance), стандартное отклонение (Std.Dev.), медиана, мода, минимальное и максимальные значения (minimum, maximum), размах (range), квантиль (quartiles), выборочный коэффициент асимметрии (skewness), выборочный коэффициент эксцесса (kurtosis).

Выборочное среднее является той точкой, сумма отклонений от которой всех рассматриваемых наблюдений равна 0. Среднее значение представляет собой характеристику положения.

Корень квадратный из выборочной дисперсии (variance) есть стандартное отклонение (Std.Dev.).

Мода - это наиболее часто встречающееся значение распределения.

Медиана - это серединное наблюдение в выборке.

Пусть имеется исходная выборка данных:

X(1) , X(2) ...., X(N)

Упорядочим их по возрастанию. Упорядоченные по возрастанию значения называют вариационным рядом:

X(1) < X(2) < ... < X(N)

Серединное значение в этом ряду называется медианой. ***X(1)*** - минимальное значение выборки, ***X(N)*** - максимальное значение выборки.

Разность между максимальным значением выборки и минимальным значением выборки называется размахом.

Асимметрия

Иногда выборочная асимметрия эксцесс используют для проверки гипотезы о том, что выборка нормальна. Для нормального распределения ***Sk=0; Ex=3***.

Корреляция есть нормированная ковариация. Коэффициент корреляции характеризует линейную зависимость между двумя случайными величинами.

Коэффициент корреляции является мерой зависимости двух величин. Коэффициент корреляции - это безразмерная величина, значение которого лежит между -1 и +1. Если при возрастании одной величины наблюдается рост другой величины, то говорят о положительной корреляции, если при возрастании одной величины наблюдается тенденция уменьшения другой величины, то говорят об отрицательной коррелированности величин.

Нулевая корреляция означает, что линейной зависимости между переменными нет. Если X,Y случайные величины, то из равенства 0 коэффициента корреляции следует независимость переменных.

В системе "Statistica" вычисляются выборочные коэффициенты корреляции и ковариации. Выборочные коэффициенты корреляции и ковариации получаются, если математические ожидания и дисперсии заменить их выборочными аналогами.

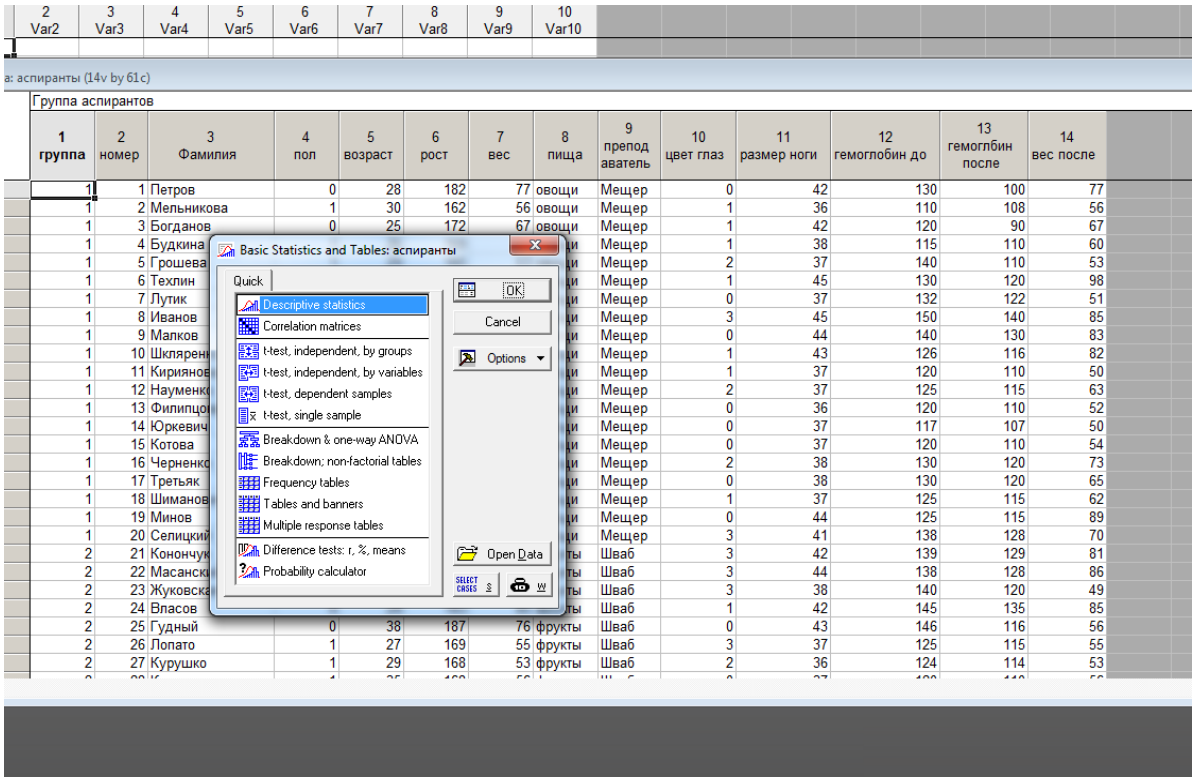

Рисунок 1. Описательная статистика

1. Открыть "Statistica".
2. Выбрать модуль "BasicStatistica".
3. Выберите в предлагаемом меню методов верхнюю строчку "DescriptiveStatistic" - "Описательные статистики.
4. "DescriptiveStatistic": рис.1.

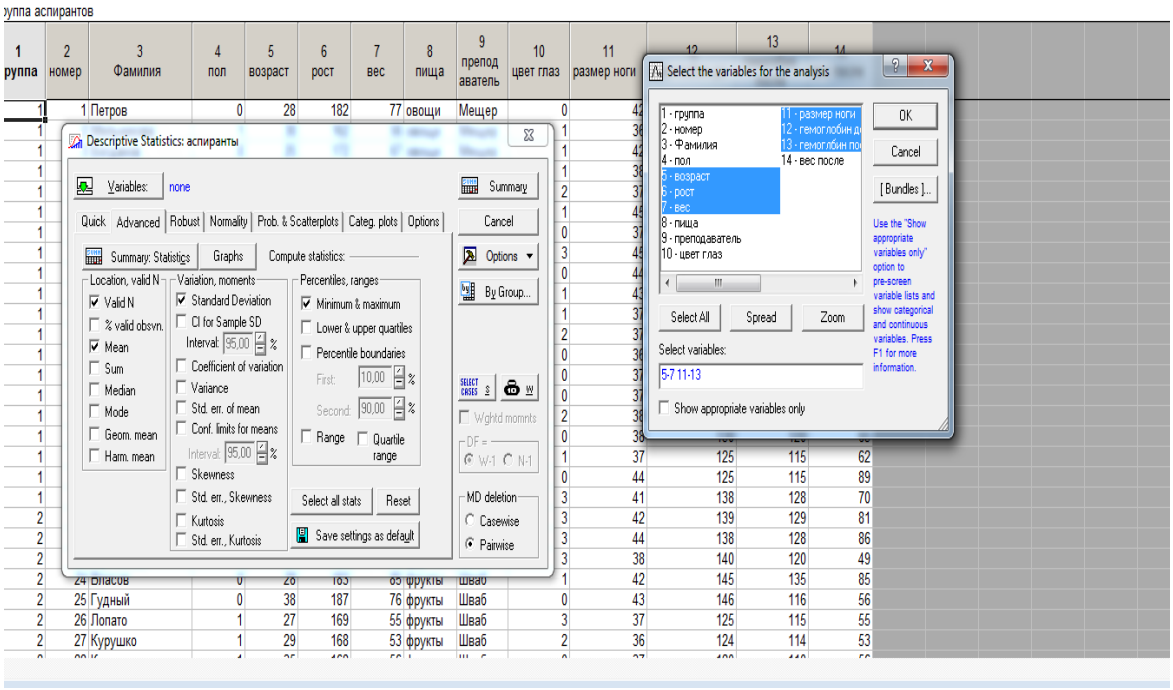

Рисунок 2. Выбор переменных.

Нажмите на кнопку "Variables" в верхней части окна и выберите для анализа переменные файла.Рис. 2.

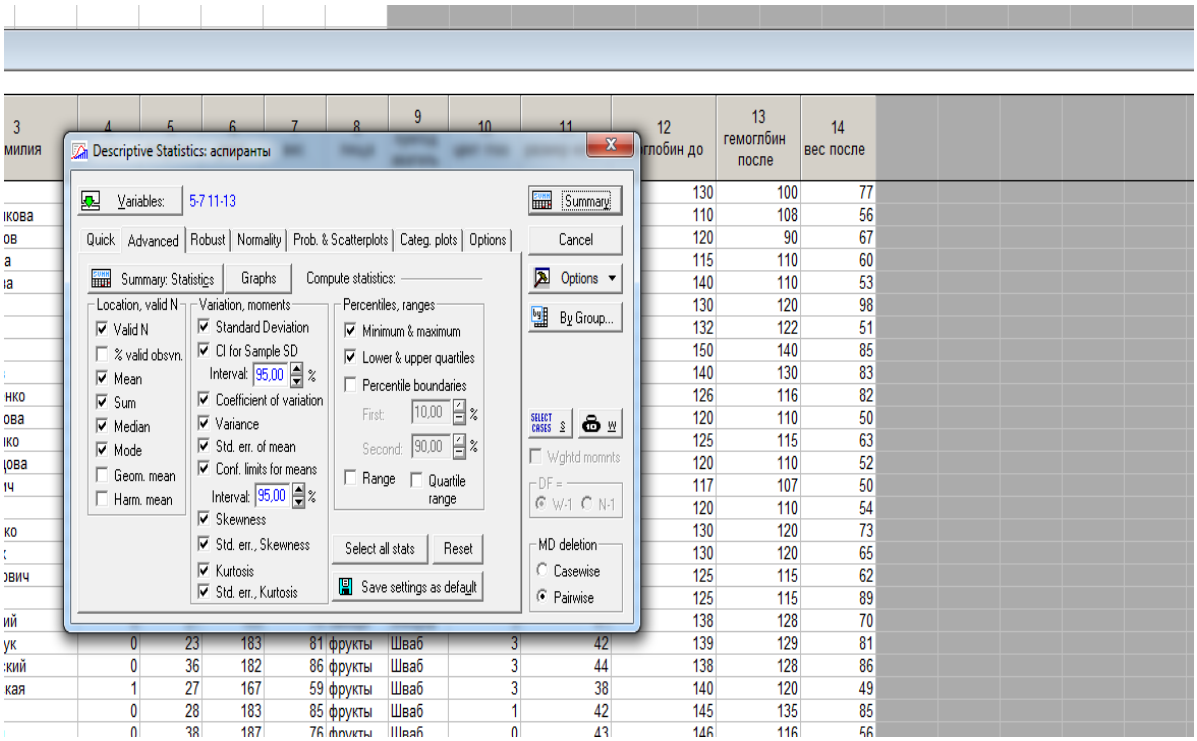

Рисунок 3. Выбор центральных моментов.

Descriptive Statistics (аспиранты)

Variable	Valid N	Mean	Confidence -95,000%	Confidence 95,000	Median	Mode	Frequency of Mode	Sum	Minimum	Maximum	Lower Quartile	Upper Quartile	Variance	Std.Dev.	Confidence SD -95,000%	Confidence SD +95,000%	Coef.Var.	Standard Error	Skewness	Std.Err. Skewness	Kurtosis	Std.Err. Kurtosis
возраст	60	28,4500	27,4530	29,4470	28,0000	27,00000	9	1707,00	22,0000	38,0000	25,5000	31,0000	14,8958	3,85950	3,27145	4,70729	13,56592	0,498260	0,642017	0,308694	-0,15468	0,608492
рост	60	173,7667	171,3241	176,2092	172,0000	184,0000	5	10426,00	152,0000	194,0000	166,0000	183,0000	89,4023	9,45528	8,01461	11,53224	5,44136	1,220671	0,085147	0,308694	-0,96845	0,608492
вес	60	74,5500	70,9802	78,1198	76,0000	89,00000	6	4473,00	50,0000	102,0000	63,5000	85,0000	190,9636	13,81896	11,71342	16,85446	18,53649	1,784020	-0,077122	0,308694	-0,94217	0,608492
размер ноги	60	40,2833	39,5031	41,0635	40,0000	37,00000	11	2417,00	36,0000	46,0000	37,0000	43,0000	9,1218	3,02022	2,56004	3,68365	7,49745	0,389909	0,142676	0,308694	-1,37425	0,608492
гемоглобин до	60	131,2167	128,3570	134,0764	130,0000	120,0000	9	7873,00	110,0000	152,0000	121,0000	139,5000	122,5455	11,07003	9,38333	13,50169	8,43645	1,429134	0,390831	0,308694	-0,84174	0,608492
гемоглбин после	60	122,2167	118,7378	125,6956	120,0000	120,0000	7	7333,00	90,0000	152,0000	114,5000	130,5000	181,3590	13,46696	11,41505	16,42514	11,01892	1,738577	0,240088	0,308694	0,41042	0,608492

Рисунок 4. Оценка данных.

Оцените близость распределения переменных к нормальному закону. Нажмите на кнопку "Histograms" . На гистограмму можно наложить плотность нормального распределения Рис.5.

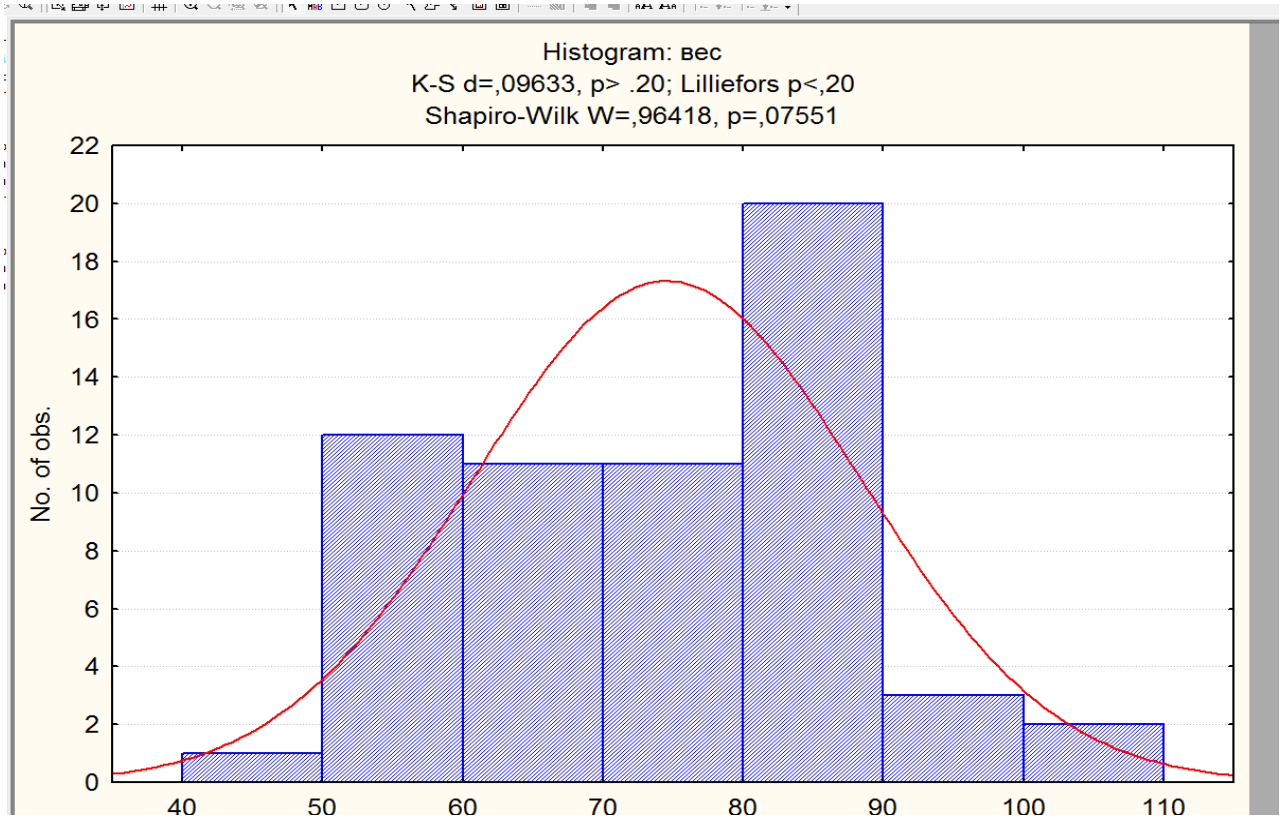

Рисунок 5. Гистограмма. Критерии согласия.

2.СТАТИСТИЧЕСКИЕ СРАВНЕНИЯ

Решение той или иной задачи не обходится, как правило, без сравнения статистических показателей, отображающих размеры и количественные соотношения анализируемых явлений. В статистике для этих целей применяется так называемая нулевая гипотеза, т. е. предположение о том, что разница между генеральными параметрами сравниваемых групп близка к нулю, а различия, которые наблюдаются между выборочными показателями, носят случайный характер. Истинность принятой гипотезы проверяется с помощью критериев значимости, т. е. случайных величин, функции распределения которых известны. Обычно для каждого критерия составляются таблицы, в которых содержатся критические величины, отвечающие определенному объему выборки и принятым уровням значимости. В биологических исследованиях принимается 5 % уровень значимости, которому отвечает нормированное отклонение t = 1,96 (2,0) при объеме выборки больше 30 единиц в случае нормального распределения признаков. Например, если окажется, что $P>0,05$, то отвергнуть нулевую гипотезу нет оснований; при $P<0,05$ нулевая гипотеза отвергается, т. е. с вероятностью более 95 % разница между выборочными показателями считается статистически значимой (достоверной).

3. ВЫПОЛНЕНИЕ МНОГОМЕРНОГО РЕГРЕССИОННОГО АНАЛИЗА В ПАКЕТЕ STATISTICA 7.0.

Медицина имеет дело прежде всего с результатами измерений, которые по природе своей представляют собой случайные величины.

Значимость регрессионной модели. Для каждого значения F можно вычислить соответствующую вероятность. Если значение этой вероятности меньше принятого уровня значимости p или вероятности ошибки (в программе Statistica это 5% или 0,05), гипотеза об отсутствии линейной связи между результативным и факторными признаками отклоняется и регрессия признается значимой.

Регрессия - история этого термина связана с исследованием зависимости признаков потомства от аналогичных признаков предков. При отсутствии отбора они постепенно *регрессируют* к общим характеристикам всего видах

$$y_x = a_0 + a_1 * x$$

Рассмотрим пример построения регрессионной модели в пакете Statistica 7.0.
Для этих целей обычно используется модуль **Multiple Regressions**(множественная регрессия), который позволяет предсказать зависимую переменную по нескольким независимым переменным.
В стартовом диалоговом окне этого модуля (рис.6) при помощи кнопки Variables указываются зависимая (dependent) и независимые(ая) (independent) переменные. В поле Inputfile указывается тип файла с данными:
RawData - данные в виде строчной таблицы;
CorrelationMatrix - данные в виде корреляционной матрицы.

Таблица 1.

4-я переменная	пол	Пол обследуемого(1 – женщина; 0 – мужчина)
5-я переменная	возраст	Возраст обследуемого, лет
6-я переменная	рост	Рост обследуемого, см
7-я переменная	вес	Вес обследуемого, кг

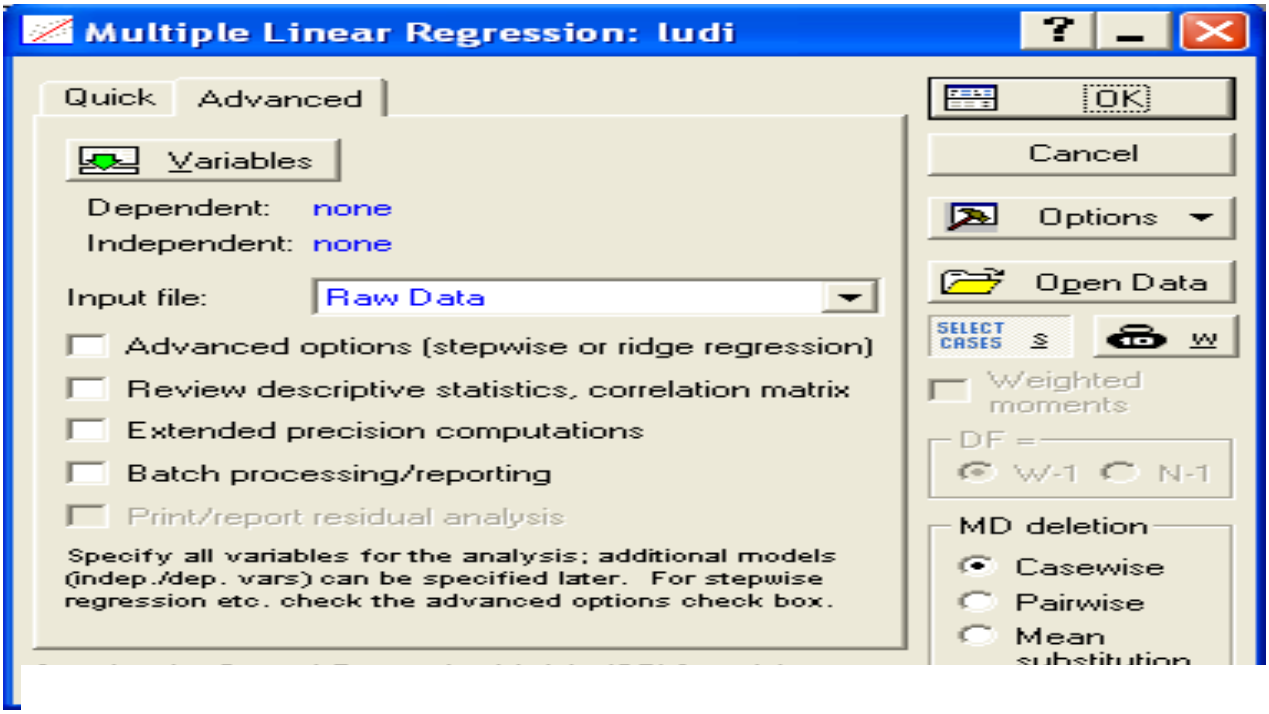

Рисунок 6. Модуль MultipleRegression.

В поле MD deletion указывается способ исключения из обработки недостающих данных:

1. **Casewise** - игнорируется вся строка, в которой есть хотя бы одно пропущенное значение;
2. M**eanSubstitution** - взамен пропущенных данных подставляются средние значения переменной;
3. **Pairwise** - попарное исключение данных с пропусками из тех переменных, корреляция которых вычисляется.

Рассмотрим проведение регрессионного анализа на конкретном примере. Имеются результаты измерения физических данных 60 аспирантов (мужчин и женщин). Из файла данных (рис.7) будем использовать 4 переменные:

	1 группа	2 номер	Фамили	4 пол	5 возраст	6 рост	7 вес	8 пища	9 преподаватель	10 цвет глаз	ра
37	2	37	Пуг	1	16	164	64	фрукты	Шваб	0	
38	2	38	Апи	1	21	162	58	фрукты	Шваб	0	
39	2	39	Май	1	29	160	59	фрукты	Шваб	0	
40	2	40	Шус	0	25	176	74	фрукты	Шваб	0	
41	3	41	Пых	0	27	183	82	мясо	Радик	0	
42	3	42	Кря	0	38	184	89	мясо	Радик	3	
43	3	43	Мал	1	21	188	102	мясо	Радик	3	
44	3	44	Гри	0	25	165	78	мясо	Радик	3	
45	3	45	Ива	0	15	178	84	мясо	Радик	3	

Рисунок 7. Окно файла данных.

Так как в файле данных содержится информация о мужчинах и женщинах, а мы хотим провести исследования только для мужчин, то воспользовавшись кнопкой Selectcases(рис. 6) можно в анализ включить только те случаи, для которых первая переменная (пол) равна 0.

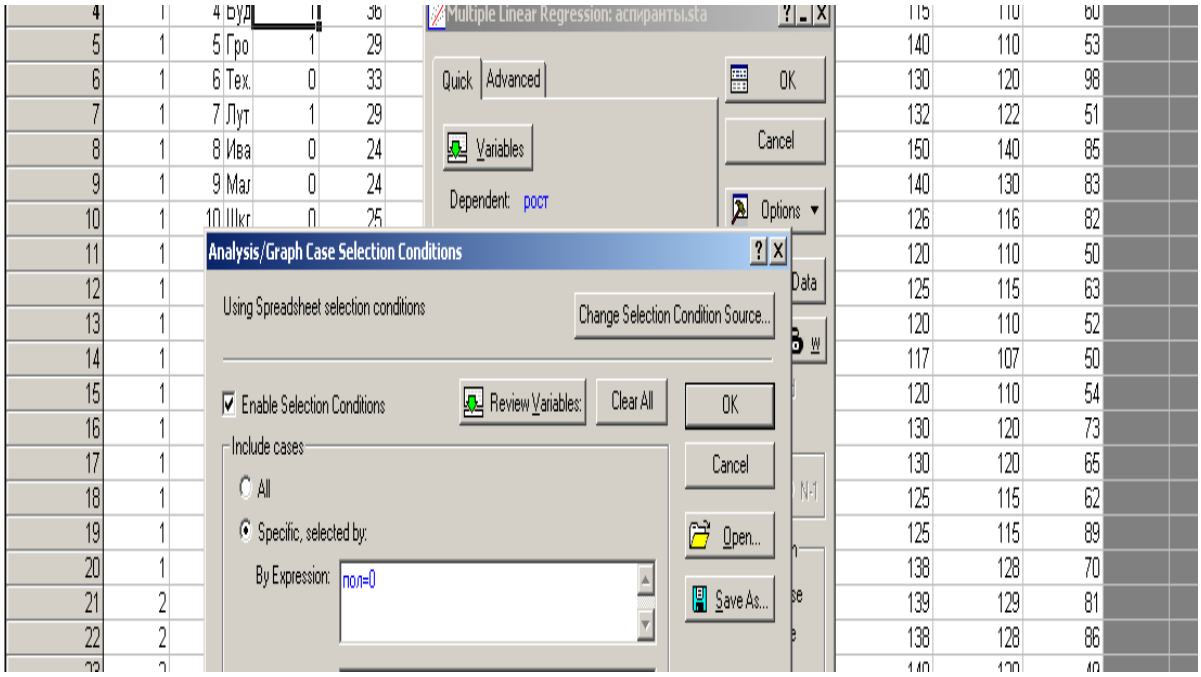

Рисунок 8. Окно включения (исключения) данных в анализ.

На первом этапе исследований учтем, что при наличии одной зависимой переменной (рост) и двух независимых переменных (возраст и вес) можно предложить различные модели линейной регрессии:

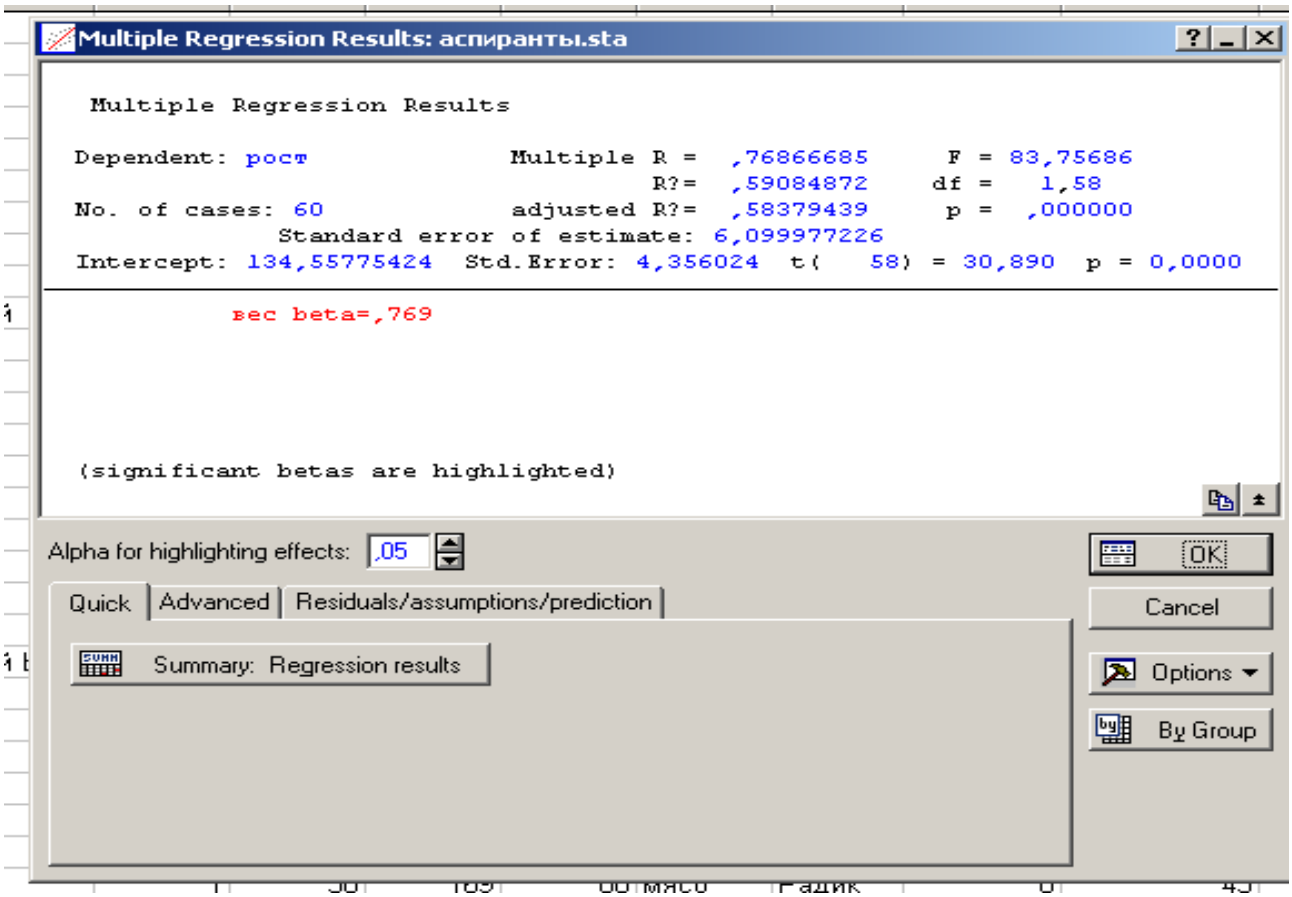

Рисунок 9.

Таблица 2

№ Модели	Вид зависимости	Комментарии
1	**рост=$a_0 + a_1$** · возраст	одномерная
2	**рост=$c_0 + c_1$** · вес	одномерная
3	**рост=$b_0 + b_1$** · возраст +· b_2 · вес	многомерная

РАССМОТРИМ МОДЕЛЬ №1

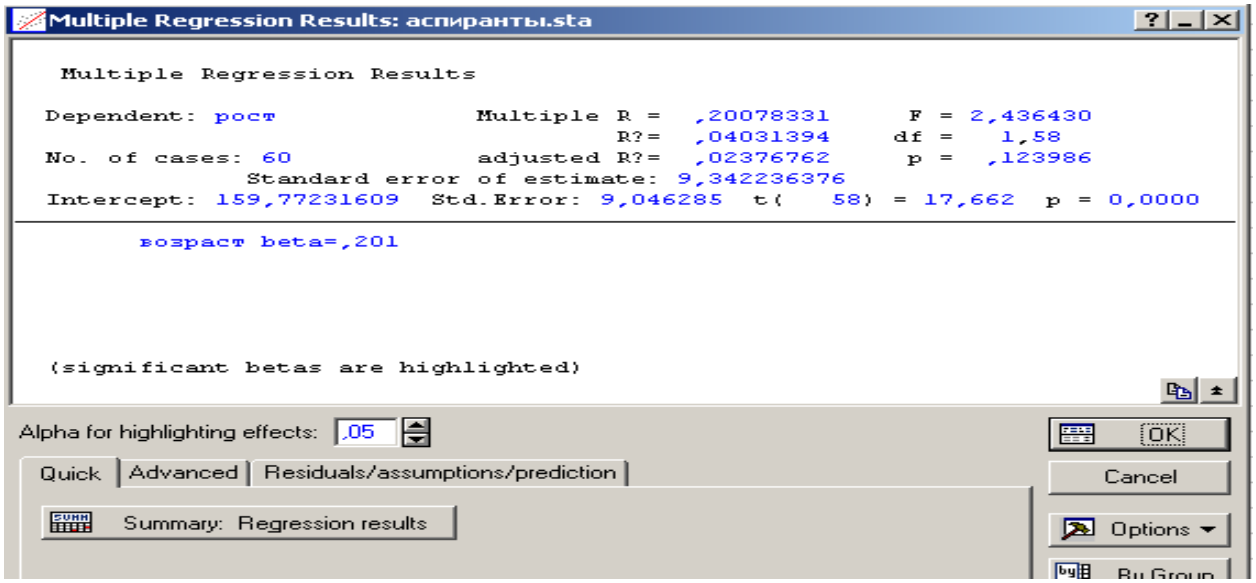

Рисунок 10.

О качестве предложенной модели регрессии будем судить по величине *коэффициента детерминации* R^2.

Модель №1 описывает 4%данных. Нулевую гипотезу H_0 принимаем, p=0,124.

Независимая переменная **возраст** незначительно влияет на зависимую**рост**.

РАССМОТРИМ МОДЕЛЬ №2

Модель №2 описывает 59%данных. Нулевую гипотезу H_0 отвергаем, p=0,000.

Независимая переменная **вес** значительно влияет на зависимую **рост**.

РАССМОТРИМ МОДЕЛЬ №3

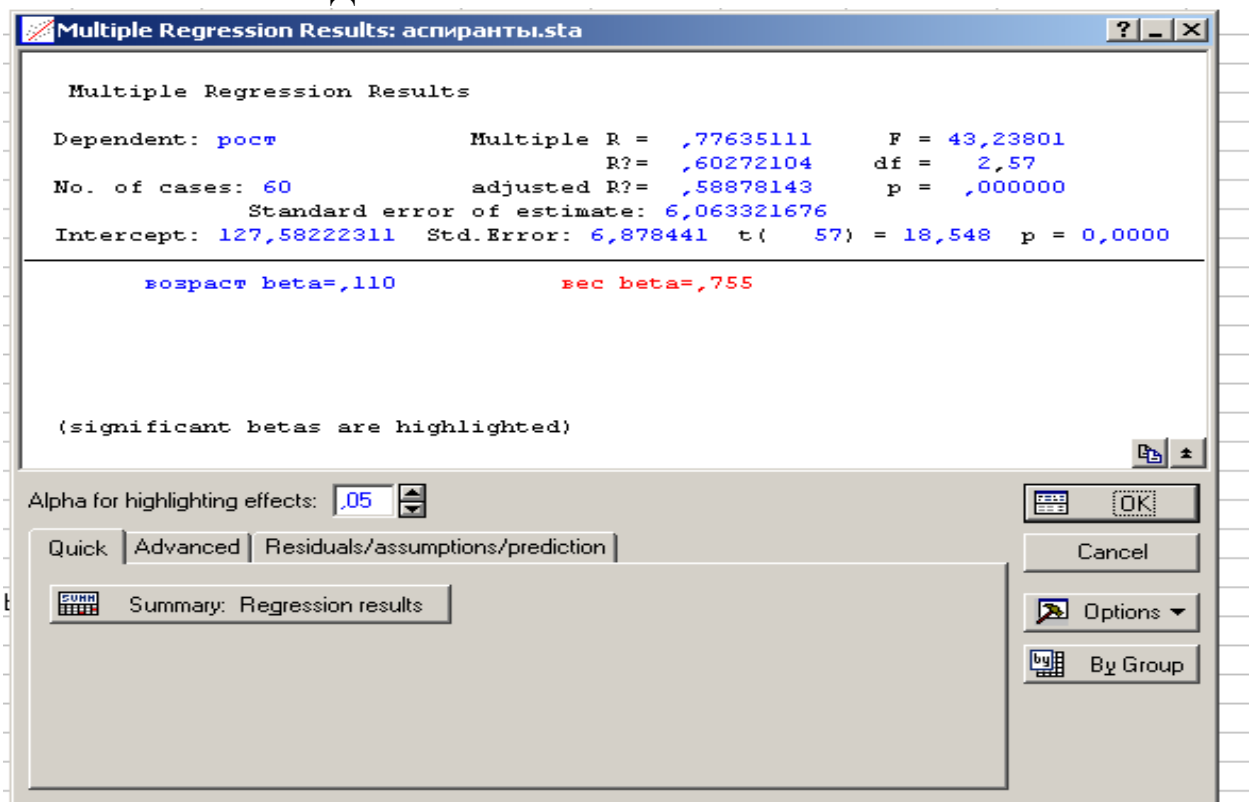

Рисунок 11. Результаты регрессионного анализа

Модель №2 описывает 60%данных. Нулевую гипотезу H_0 отвергаем, p=0,000.

Независимые переменные **вес** и **возраст** значительно влияют на зависимую **рост**.

Если в качестве критерия оптимизации выбрать простоту модели (одномерная) и если добавить ещё один критерий – максимальный % описания данных, то из этих двух моделей выбираем модель №2. Теперь в качестве главного критерия оптимизации выбираем максимальный процент описания данных и сравниваем модели №2 и №3. Нужно сказать, что модель №3 – многомерная, а модель №2 – одномерная. Таким образом, на первом этапе можно сказать, что многомерная модель №3 более адекватна и лучше описывает исходные данные. Естественно предположить, что и предсказания по модели №3 будут более надежными (точными).

Теперь более подробно рассмотрим последовательность действий создания модели и анализ полученных результатов.

После выбора всех опций стартового диалогового окна регрессионного анализа и нажатия кнопки **ОК** появляется окно результатов регрессионного анализа MultipleRegressionsResults (см. рис. 9).Детально проанализируем полученные результаты регрессионной модели.

В верхней части окна приведены наиболее важные параметры полученной регрессионной модели:

1. **Multiple R** - коэффициент множественной корреляции, который характеризует тесноту линейной связи между зависимой и всеми независимыми переменными. Может принимать значения от 0 до 1.
2. $\mathbf{R^2}$ **-** коэффициент детерминации. Численно выражает долю вариации зависимой переменной, объясненную с помощью регрессионного уравнения. Чем больше $\mathbf{R^2}$, тем большую долю вариации объясняют переменные, включенные в модель.Коэффициент множественной корреляции *R* является обобщением коэффициента парной корреляции для случая, когда число независимых факторов, включенных в уравнение, больше одного. *R* является величиной безразмерной. *R* не меняется при изменении единиц измерения соответствующих признаков. *R* принимает значения в интервале [0;1]. •Чем больше *R* , тем сильнее линейная связь между совокупностью независимых факторов и результативным признаком. Как и в случае парной зависимости, интерпретируется не сам коэффициент корреляции, а его квадрат – коэффициент детерминации. Этот коэффициент является квадратом соответствующего коэффициента корреляции и выражается в процентах. Смысл коэффициента детерминации R^2Коэффициент детерминации R^2 показывает, *насколько изменения зависимого признака (в процентах) объясняются изменениями совокупности независимых признаков*. То есть, это доля дисперсии зависимого признака, объясняемая влиянием независимых признаков.

3. **adjusted R** - скорректированный коэффициент множественной корреляции. Включение новой переменной в регрессионное уравнение увеличивает $\mathbf{R^2}$ не

всегда, а только в том случае, когда частный **F**-критерий при проверке гипотезы о значимости включаемой переменной больше или равен 1. В противном случае включение новой переменной уменьшает значение $\mathbf{R^2}$ и **adjusted R**.

- **F** - F-критерий используется для проверки значимости регрессии. В данном случае в качестве нулевой гипотезы проверяется гипотеза: между зависимой и независимыми переменными нет линейной зависимости;
- **df** - числа степеней свободы для F-критерия;
- **p** - вероятность нулевой гипотезы для F-критерия;
- **Standarderrorofestimate** - стандартная ошибка оценки (уравнения); Эта оценка является мерой рассеяния наблюденных значений относительно регрессионной прямой;
- **Intercept** – оценка свободного члена уравнения;
- **Std.Error** - стандартная ошибка оценки свободного члена уравнения;
- **t** - t-критерий для оценки свободного члена уравнения;
- **p** - вероятность нулевой гипотезы для свободного члена уравнения.
- **Beta** - β-коэффициенты уравнения. Это стандартизированные регрессионные коэффициенты, рассчитанные по стандартизированным значениям переменных. По их величине можно оценить значимость зависимых переменных. Коэффициент показывает, на сколько единиц стандартного отклонения изменится зависимая переменная при изменении на одно стандартное отклонение независимой переменной, при условии постоянства остальных независимых переменных. Свободный член в таком уравнении равен 0.

Нажатие кнопки 

- в окне результатов (см рис. 10) позволяет получить основные результаты регрессионной модели (рис. 11), часть из которых уже была описана: **B** - коэффициенты уравнения регрессии; **St. Err. of B** - стандартные ошибки коэффициентов уравнения регрессии; **t** (11) - t-критерий для коэффициентов уравнения регрессии; **p-level** - вероятность нулевой гипотезы для коэффициентов уравнения регрессии.

Таблица 3

Regression Summary for Dependent Variable: рост (аспиранты.sta) R= ,776 R^2= ,602 Adjusted R^2= ,588 F(2,57)=43,238 p

	Beta	Std.Err. - ofBeta	B	Std.Err. - of B	t(57)	p-level
Intercept			127,5822	6,878441	18,54813	0,000000
возраст	0,109760	0,084098	0,2689	0,206029	1,30514	0,197089
вес	0,755442	0,084098	0,5169	0,057542	8,98287	0,000000

Параметры уравнения регрессии

В результате проведенного анализа было получено следующее уравнение:

рост = 127,582 + 0,268*возраст + 0,516*вес.

Это уравнение объясняет 60,27% ($\mathbf{R^2 =, 6027}$) вариации зависимой переменной. Полученные результаты свидетельствуют о том что коэффициент $\boldsymbol{b_2}$ при переменной **возраст** незначимо отличается от нуля, однако включение этой переменной в регрессионную модель увеличивает на 1% процент исходных данных, корректно описанных регрессионным уравнением.

Проверка качества уравнения регрессии осуществлялась с помощью статистики $F = 43{,}238$ По статистическим таблицам Фишера – Снедекора с данными степенями свободы $(\mathrm{df} = 2{,}57)$гипотезуH_0 (линейная зависимость отсутствует) можно принять с вероятностью$(p = 0.00000)$; при уровне значимости α = 0.05 принимаем альтернативную гипотезу – линейная зависимость значима.

Одновременно проверялась статистическая значимость коэффициентов множественной регрессии (критерий Стьюдента). Видно (см. рис. 5), что коэффициенты $\boldsymbol{b_0}$ и $\boldsymbol{b_2}$ значимо отличаются от нуля, коэффициент $\boldsymbol{b_1}$ незначимо отличается от нуля.

Для расчета по полученному регрессионному уравнению значений зависимой переменной по значениям независимых переменных воспользуемся кнопкой 

(раздел Residuals/assumptions/prediction) (рис.10).

Зададим значения возраста (возраст = 55) и веса (вес = 85). Учтем, что в пакете Statistica 7.0. приводится как точечная, так и интервальная оценка (рис. 12).

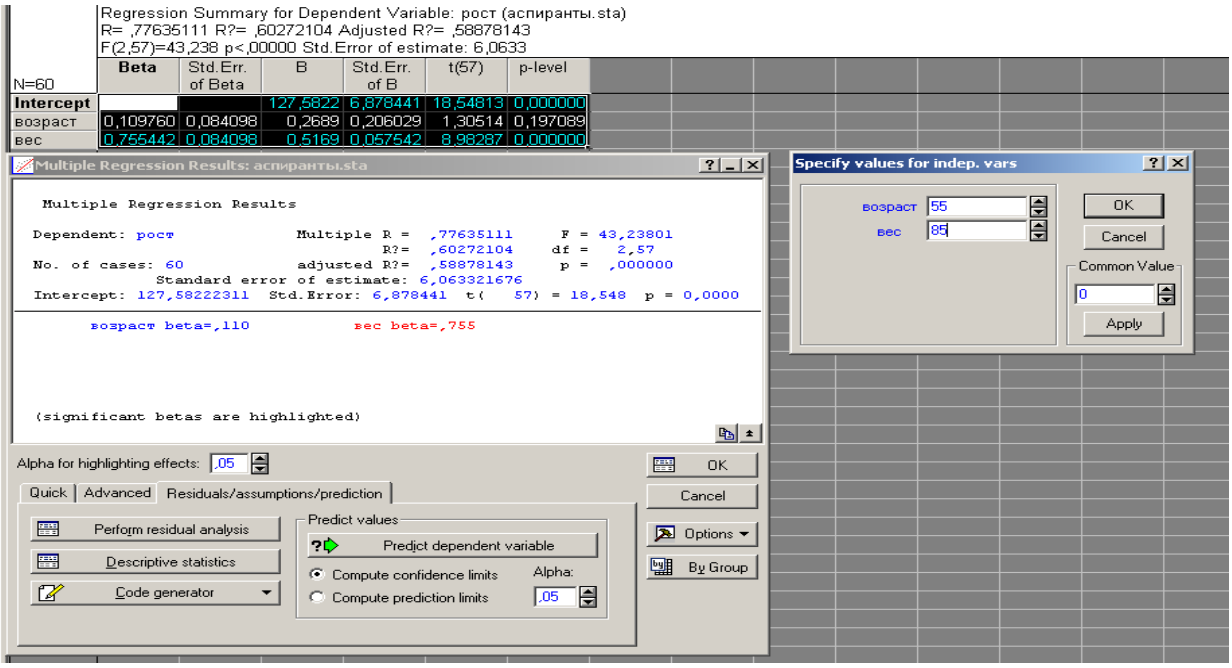

Рисунок 12. Окно задание значений независимых переменных

Variable	Predicting Values for (аспиранты.sta) variable: рост		
	B-Weight	Value	B-Weight * Value
возраст	0,268897	55,00000	14,7893
вес	0,516892	85,00000	43,9358
Intercept			127,5822
Predicted			186,3074
-95,0%CL			175,3206
+95,0%CL			197,2943

Рисунок 13. Предсказанные точечные и интервальные значения

О полученных результатах можно сказать следующее: рост = 186,307– это точечная оценка. 95% доверительный интервал равен (175.307;197,294).

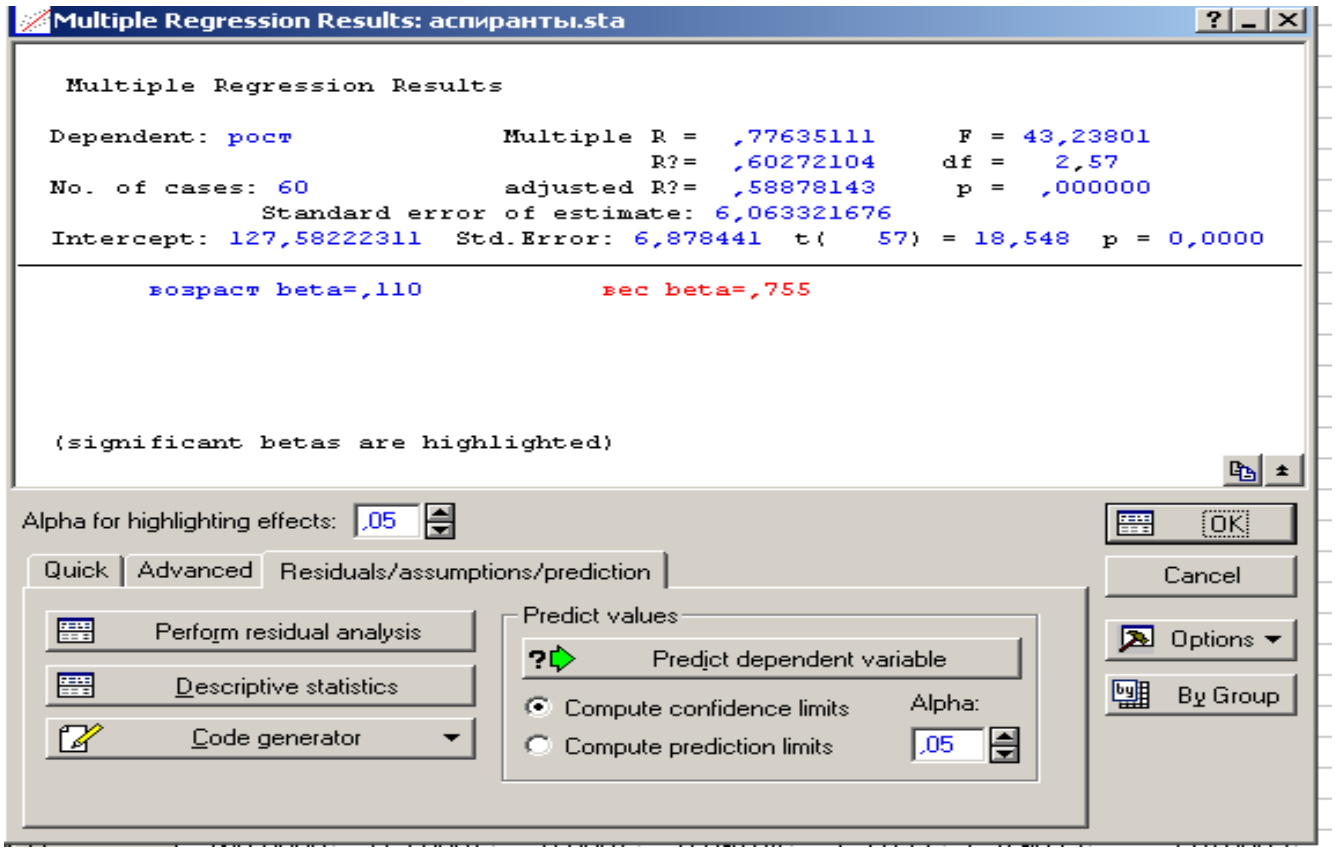

Рисунок 14. Оценка величины остатков

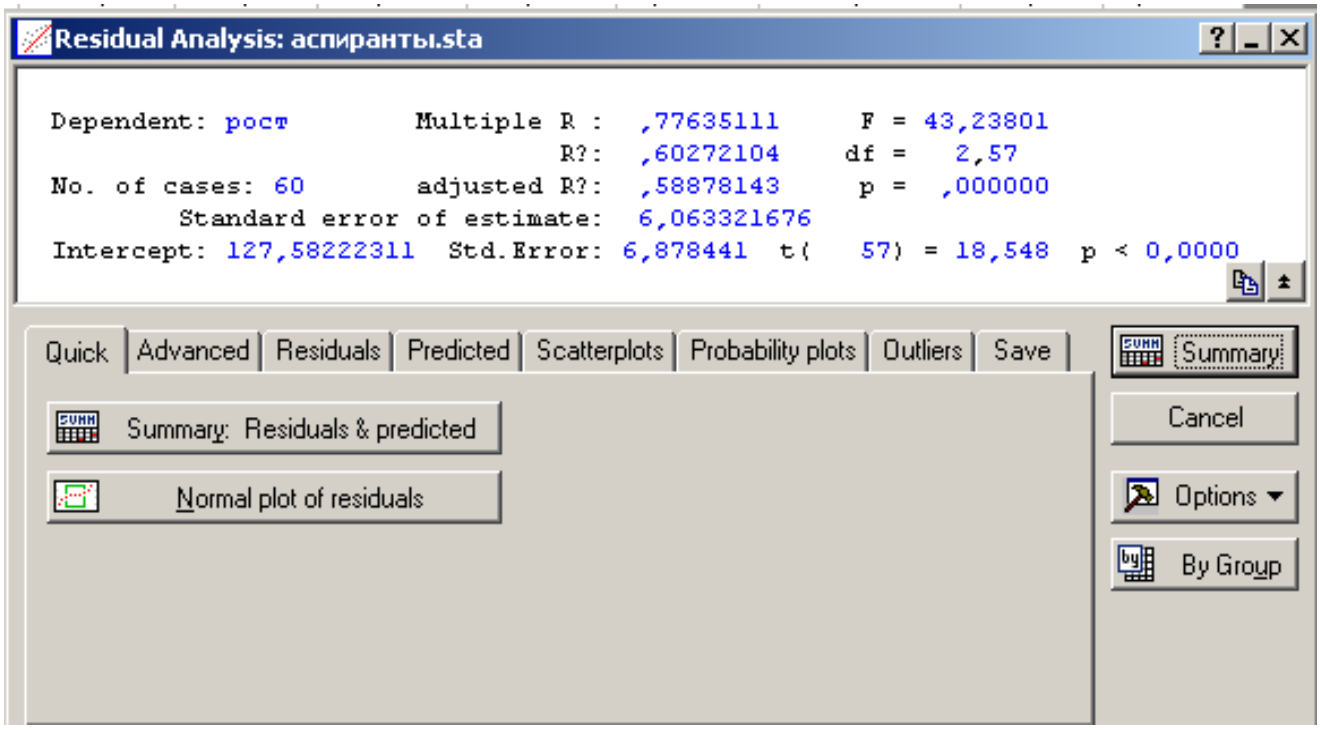

Рисунок 15. Специальные критерии

При нажатии на кнопку 

можно оценить величины остатков и специальных критериев (см. рис. 12и рис. 13.).

В таблицу включены все случаи, приведены исходные данные (Observed), данные модели (Predicted) и остатки (Residual). Остатки – это разность исходных и предсказанных данных.

Таблица4

Predicted & Residual Values (аспиранты.sta) Dependent variable: рост

	Observed - Value	Predicted - Value	Residual	Standard - Pred. v.	Standard - Residual	Std.Err. - Pred.Val	Mahalanobis - Distance	Deleted - Residual	Cook's - Distance
1	182,0000	174,9120	7,0880	0,15603	1,16899	0,802715	0,050745	7,2144	0,008271
2	162,0000	164,5951	-2,5951	-1,24943	-0,42800	1,391477	2,123963	-2,7394	0,003583
3	172,0000	168,9364	3,0636	-0,65801	0,50526	1,110087	0,994297	3,1698	0,003054
4	169,0000	165,5871	3,4129	-1,11429	0,56288	1,211053	1,370397	3,5547	0,004571
5	165,0000	162,7755	2,2245	-1,49730	0,36687	1,482265	2,542666	2,3659	0,003033
6	194,0000	187,1113	6,8887	1,81791	1,13613	1,734203	3,843147	7,5025	0,041749
7	164,0000	161,7417	2,2583	-1,63813	0,37245	1,580794	3,027009	2,4229	0,003618
8	184,0000	177,9716	6,0284	0,57283	0,99424	1,395615	2,142472	6,3656	0,019465
9	184,0000	176,9378	7,0622	0,43200	1,16474	1,340580	1,900807	7,4251	0,024436
10	180,0000	176,6898	3,3102	0,39822	0,54594	1,172677	1,223593	3,4388	0,004011
11	164,0000	160,9560	3,0440	-1,74518	0,50204	1,607906	3,165751	3,2743	0,006836
12	160,0000	168,2133	-8,2133	-0,75652	-1,35460	1,098893	0,954611	-8,4923	0,021478
13	152,0000	160,9142	-8,9142	-1,75088	-1,47018	1,688289	3,590962	-9,6634	0,065643
14	170,0000	171,0249	-1,0249	-0,37351	-0,16903	0,866993	0,222986	-1,0463	0,000203
15	163,0000	169,5369	-6,5369	-0,57621	-1,07811	1,413217	2,221818	-6,9125	0,023535
16	166,0000	174,4579	-8,4579	0,09416	-1,39492	1,397415	2,150541	-8,9323	0,038425
17	166,0000	168,4404	-2,4404	-0,72558	-0,40249	0,982037	0,564365	-2,5062	0,001494
18	166,0000	165,5453	0,4547	-1,11999	0,07499	1,633683	3,299847	0,4903	0,000158
19	185,0000	182,1903	2,8097	1,14754	0,46339	1,300958	1,732841	2,9453	0,003621
20	169,0000	171,0249	-2,0249	-0,37351	-0,33396	0,866993	0,222986	-2,0672	0,000792
21	183,0000	175,6351	7,3649	0,25454	1,21466	1,453176	2,405634	7,8137	0,031797

22	182,0000	181,7153	0,2847	1,08282	0,04696	1,794295	4,183429	0,3121	0,000077
23	167,0000	165,3391	1,6609	-1,14808	0,27393	1,199246	1,324727	1,7285	0,001060
24	183,0000	179,0472	3,9528	0,71936	0,65192	0,998168	0,615629	4,0629	0,004056
25	187,0000	177,0841	9,9159	0,45193	1,63539	2,109858	6,160596	11,2819	0,139737
26	169,0000	163,2715	5,7285	-1,42974	0,94478	1,373498	2,044184	6,0383	0,016964
27	168,0000	162,7755	5,2245	-1,49730	0,86165	1,482265	2,542666	5,5565	0,016730
28	168,0000	165,9396	2,0604	-1,06627	0,33982	1,979983	5,308158	2,3063	0,005143
29	172,0000	162,7337	9,2663	-1,50300	1,52825	1,480129	2,532513	9,8534	0,052458
30	186,0000	177,5174	8,4826	0,51096	1,39900	0,907689	0,338891	8,6770	0,015299
31	188,0000	177,9716	10,0284	0,57283	1,65394	1,395615	2,142472	10,5894	0,053866
32	165,0000	173,0507	-8,0507	-0,09754	-1,32777	1,379523	2,070803	-8,4902	0,033832
33	178,0000	180,8459	-2,8459	0,96439	-0,46936	1,205487	1,348815	-2,9630	0,003146
34	190,0000	188,9099	1,0901	2,06294	0,17978	1,834174	4,415647	1,1999	0,001194
35	166,0000	171,0458	-5,0458	-0,37066	-0,83219	0,858069	0,198279	-5,1489	0,004814
36	172,0000	174,9539	-2,9539	0,16173	-0,48717	1,069481	0,852262	-3,0487	0,002622
37	164,0000	167,6546	-3,6546	-0,83263	-0,60275	1,078060	0,881829	-3,7740	0,004082
38	162,0000	165,8978	-3,8978	-1,07197	-0,64285	1,384280	2,091906	-4,1121	0,007991
39	160,0000	165,8769	-5,8769	-1,07481	-0,96925	1,204422	1,344696	-6,1183	0,013392
40	176,0000	172,5547	3,4453	-0,16511	0,56822	1,055249	0,803732	3,5529	0,003467
41	183,0000	177,2276	5,7724	0,47148	0,95202	0,957399	0,487681	5,9200	0,007923
42	184,0000	183,8037	0,1963	1,36733	0,03237	2,186599	6,689736	0,2256	0,000060
43	188,0000	188,9099	-0,9099	2,06294	-0,15007	1,834174	4,415647	-1,0016	0,000832
44	165,0000	174,6223	-9,6223	0,11655	-1,58696	1,091505	0,928643	-9,9445	0,029057
45	178,0000	180,4126	-2,4126	0,90536	-0,39790	1,597721	3,113351	-2,5926	0,004232
46	174,0000	181,6317	-	1,07143	-1,25866	1,196475	1,314075	-7,9409	0,022263

			7,6317						
47	184,0000	174,2099	9,7901	0,06038	1,61465	1,582121	3,033743	10,5054	0,068130
48	172,0000	180,5770	-8,5770	0,92776	-1,41457	1,288421	1,680743	-8,9826	0,033033
49	164,0000	168,1716	-4,1716	-0,76221	-0,68800	1,050076	0,786256	-4,3005	0,005029
50	162,0000	174,1263	-12,1263	0,04899	-1,99994	0,845545	0,164039	-12,3668	0,026966
51	180,0000	179,7703	0,2297	0,81786	0,03789	1,670298	3,493992	0,2486	0,000043
52	176,0000	181,0730	-5,0730	0,99532	-0,83666	1,599509	3,122526	-5,4524	0,018758
53	183,0000	179,7912	3,2088	0,82071	0,52922	1,359971	1,984844	3,3788	0,005207
54	184,0000	177,4756	6,5244	0,50526	1,07604	1,078467	0,883236	6,7376	0,013021
55	179,0000	173,0925	5,9075	-0,09184	0,97431	0,837079	0,141178	6,0223	0,006268
56	170,0000	170,2600	-0,2600	-0,47770	-0,04288	0,868884	0,228254	-0,2655	0,000013
57	168,0000	181,3837	-13,3836	1,03765	-2,20731	1,137642	1,093692	-13,8720	0,061422
58	169,0000	177,0005	-8,0005	0,44054	-1,31949	0,888215	0,282764	-8,1760	0,013006
59	179,0000	178,3032	0,6968	0,61800	0,11492	1,009083	0,650790	0,7166	0,000129
60	181,0000	184,7748	-3,7748	1,49962	-0,62256	1,484254	2,552139	-4,0154	0,008760
Minimum	152,0000	160,9142	-13,3836	-1,75088	-2,20731	0,802715	0,050745	-13,8720	0,000013
Maximum	194,0000	188,9099	10,0284	2,06294	1,65394	2,186599	6,689736	11,2819	0,139737
Mean	173,7667	173,7667	-0,0000	0,00000	-0,00000	1,314733	1,966667	0,0329	0,016898
Median	172,0000	174,5401	0,3697	0,10536	0,06098	1,320769	1,816824	0,4012	0,006552

Таблица остатков.

Для выделения имеющихся в регрессионных остатках выбросов предложен ряд дополнительных показателей:

- Расстояние Кука (Cook'sDistance) – принимает только положительное значение и показывает расстояние между коэффициентами уравнения регрессии после исключения из обработки i-ой точки данных. Большое значение показателя Кука указывает на сильно влияющий случай (выброс).

В нашем случае Case № 25 смещает оценки коэффициентов регрессии.

- Расстояние Махаланобиса (Mahalns.Distance) – показывает насколько каждый случай или точка в p-мерном пространстве независимых переменных отклоняется от центра статистической совокупности.

Кнопка 

(раздел Advanced) предназначена для поиска выбросов. Выбросы – это остатки, которые значительно превосходят по абсолютной величине остальные. Выбросы показывают опытные данные, которые являются не типичными по отношению к остальным данным, и требует выяснения причин их возникновения. Выбросы должны исключаться из обработки, если они вызваны ошибками регистрации, измерения и т.п.

ЛИТЕРАТУРА

1. О.Ю.Реброва. Статистический анализ медицинских данных. Применение пакета прикладных программ Statistika.– М. :МЕДИАСФЕРА, 2002.312 с.
2. А.Петри, К.Сэбин. Наглядная статистика в медицине.– М. :Издательский дом, 2005.162 с.
3. С.Гланц. Медико-биологическая статистика. Пер. с англ.-М.Практика, 1998.-459 с.
4. Д.Худсон. Статистика для физиков.-М.: МИР ,1970. 296 с.
5. А.А.Халафян . STATISTICA 6. Статистический анализ данных. 3-е изд. Учебник – М.:ООО «Бином-Пресс», 2007 г.-512 с.:ил.
6. Боровиков В. Statistica. Искусство анализа данных на компьютере. 2-ое изд. – СПб.: Питер,2003 г.- 688 стр.
7. Т.Гринсальх. Основы доказательной медицины: Пер.с англ. – М.:ГЭОТАР-Медиа,2006.-240с.:ил.
8. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. – М.: Наука, 1983. – 416 с.
9. Зайцев В.М., Лифляндский В.Г., Маринкин В.И. Прикладная медицинская статистика: Учебное пособие. СПб, 2003.
10. Информатика для медиков [Электронный ресурс]: учебное пособие / Г. А. Хай. - СПб. : СпецЛит, 2009. - 223 с. – Режим доступа: http://www.studmedlib.ru
11. Ланг Т.А. Как описывать статистику в медицине. Аннотированное руководство для авторов, редакторов и рецензентов / Т.А. Ланг, М Сесик; пер. с англ. Под ред. В. П. Леонова. – М.: Прикладная медицина, 2011. – 480 с.: ил.
12. Основы высшей математики и математической статистики [Электронный ресурс]: учебник / И.В. Павлушков и др.: 2-е изд., испр. - М. : ГЭОТАР-Медиа, 2009. - 432 с.: ил. – Режим доступа: http://www.studmedlib.ru
13. Статистические методы анализа в здравоохранении. Краткий курс лекций. [Электронный ресурс]: Подготовлены авторским коллективом в

составе: д.м.н., проф. Леонов С.А., при участии к.м.н. Вайсман Д.Ш., Моравская С.В, Мирсков Ю.А. - М.: ИД "Менеджер здравоохранения", 2011. - 172 с. – Режим доступа: http://www.studmedlib.ru

14. О.П.Минцер, Ю.В. Вороненко, В.В.Власов. Информационные технологии в охране здоровья и практической медицине.– К. :Высшая школа, 2003.350 с.:ил.

Приложение 1

Распределение Фишера.

Значения квантилей для степеней свободы f_1 и f_2 и вероятности

λ=0,05

$f_2 \backslash f_1$	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
2	18.512	18.999	19.163	19.248	19.298	19.329	19.371	19.414	19.453
3	10.129	9.552	9.276	9.118	9.014	8.941	8.844	8.744	8.638
4	7.710	6.945	6.591	6.388	6.257	6.164	6.041	5.912	5.774
5	6.607	5.786	5.410	5.192	5.050	4.950	4.818	4.678	4.527
6	5.987	5.143	4.756	4.388	4.284	4.147	4.000	3.841	3.669
7	5.591	4.737	4.347	4.121	3.972	3.866	3.725	3.574	3.410
8	5.317	4.459	4.067	3.838	3.688	3.580	3.438	3.284	3.116
9	5.117	4.256	3.863	3.633	3.482	3.374	3.230	3.073	2.900
10	4.965	4.103	3.708	3.478	3.326	3.217	3.072	2.913	2.737
11	4.844	3.982	3.587	3.357	3.204	3.094	2.948	2.778	2.609
12	4.747	3.885	3.490	3.259	3.106	2.999	2.848	2.686	2.505
13	4.667	3.805	3.410	3.179	3.025	2.915	2.767	2.604	2.420
14	4.600	3.739	3.344	3.112	2.958	2.848	2.699	2.534	2.349
15	4.543	3.683	3.287	3.056	2.901	2.790	2.641	2.475	2.288
16	4.494	3.634	3.239	3.007	2.853	2.741	2.591	2.424	2.235
17	4.451	3.592	3.197	2.965	2.810	2.699	2.548	2.381	2.190
18	4.414	3.555	3.160	2.928	2.773	2.661	2.510	2.342	2.150
19	4.381	3.522	3.127	2.895	2.740	2.629	2.477	2.308	2.114
20	4.351	3.493	3.098	2.866	2.711	2.599	2.447	2.278	2.083
21	4.325	3.467	3.072	2.840	2.685	2.573	2.421	2.250	2.054
22	4.301	3.443	3.049	2.817	2.661	2.549	2.397	2.226	2.028
23	4.279	3.422	3.028	2.795	2.640	2.528	2.375	2.203	2.005
24	4.260	3.403	3.009	2.777	2.621	2.508	2.355	2.183	1.984
25	4.242	3.385	2.991	2.759	2.603	2.490	2.337	2.165	1.965
26	4.225	3.369	2.975	2.743	2.587	2.474	2.321	2.148	1.947
27	4.210	3.354	2.961	2.728	2.572	2.459	2.305	2.132	1.930
28	4.196	3.340	2.947	2.714	2.558	2.445	2.292	2.118	1.915
29	4.183	3.328	2.934	2.702	2.545	2.432	2.278	2.104	1.901
30	4.171	3.316	2.922	2.690	2.534	2.421	2.266	2.092	1.887
60	4.001	3.151	2.758	2.525	2.368	2.254	2.097	1,918	1.700
120	3.920	3.072	2.680	2.447	2.290	2.175	2.016	1.834	1.608

Приложение 2

Распределение Стьюдента.

Значения квантилей для степеней свободы f и вероятности

λ=0,05

f	1	2	3	4	5	6	7	8	9
t	12.71	4.303	3.182	2.776	2.571	2.447	2.365	2.306	2.262
f	10	11	12	13	14	15	16	17	18
t	2.228	2.201	2.179	2.160	2.145	2.131	2.120	2.110	2.101
f	19	20	21	22	23	24	25	26	27
t	2.093	2.086	2.080	2.074	2.069	2.064	2.060	2.056	2.052
f	28	29	30	40	50	60	80	100	200
t	2.048	2.045	2.042	2.021	2.009	2.000	1.990	1.984	1.972

Приложение 3

Критерий Дарбина-Уотсона.

Нижние и верхние границы критерия для вероятности

λ=0,05

F_1	f_2=1	f_2=2	f_2=3	f_2=4	f_2=5	f_2=6	f_2=7
	*d*н*d*в	*d*н*d*в	*d*н*d*в	*d*н*d*в	*d*н*d*в	*d*н*d*в	*d*н*d*в
6	0.61 1.40						
7	0.70 1.36	0.47 1.90					
8	0.76 1.33	0.56 1.78	0.37 2.29				
9	0.82 1.32	0.63 1.70	0.46 2.13	0.30 2.59			
10	0.88 1.32	0.70 1.64	1.53 2.02	0.38 2.41	0.24 2.82		
11	0.93 1.32	0.76 1.60	0.60 1.93	0.44 2.28	0.32 2.65	0.20 3.01	
12	0.97 1.33	0.81 1.58	0.66 1.86	0.51 2.18	0.38 2.51	0.27 2.83	0.17 3.15
13	1.01 1.34	0.86 1.56	0.72 1.82	0.57 2.09	0.45 2.39	0.33 2.69	0.23 2.99
14	1.05 1.35	0.91 1.55	0.77 1.78	0.63 2.03	0.51 2.30	0.39 2.57	0.29 2.85
15	1.08 1.36	0.95 1.54	0.81 1.75	0.69 1.98	0.56 2.22	0.45 2.47	0.34 2.73
16	1.11 1.37	0.98 1.54	0.86 1.73	0.73 1.94	0.62 2.16	0.50 2.39	0.40 2.62
17	1.13 1.38	1.02 1.54	0.90 1.71	0.78 1.90	0.66 2.10	0.55 2.32	0.45 2.54
18	1.16 1.39	1.05 1.54	0.93 1.70	0.82 1.87	0.71 2.06	0.60 2.26	0.50 2.46
19	1.18 1.40	1.07 1.54	0.97 1.69	0.86 1.85	0.75 2.02	0.65 2.21	0.55 2.40
20	1.20 1.41	1.10 1.54	1.00 1.68	0.89 1.83	0.79 1.99	0.69 2.16	0.60 2.34
25	1.29 1.45	1.21 1.55	1.12 1.65	1.04 1.77	0.95 1.89	0.87 2.01	0.78 2.14
30	1.35 1.49	1.28 1.57	1.21 1.65	1.14 1.74	1.07 1.83	1.00 1.93	0.93 2.03
40	1.44 1.54	1.39 1.60	1.34 1.66	1.29 1.72	1.23 1.79	1.18 1.85	1.12 1.92
50	1.50 1.59	1.46 1.63	1.42 1.67	1.38 1.72	1.34 1.77	1.29 1.82	1.25 1.88
60	1.55 1.62	1.51 1.65	1.48 1.69	1.44 1.73	1.41 1.77	1.37 1.81	1.34 1.85
70	1.58 1.64	1.55 1.67	1.53 1.70	1.49 1.74	1.46 1.77	1.43 1.80	1.40 1.84
80	1.61 1.66	1.59 1.69	1.56 1.72	1.53 1.74	1.51 1.77	1.48 1.80	1.45 1.83
90	1.64 1.68	1.61 1.69	1.59 1.73	1.57 1.75	1.54 1.78	1.52 1.80	1.49 1.83
100	1.65 1.69	1.63 1.70	1.61 1.74	1.59 1.76	1.57 1.78	1.55 1.80	1.53 1.83
150	1.72 1.75	1.71 1.76	1.69 1.77	1.68 1.79	1.67 1.80	1.65 1.82	1.64 1.83
200	1.76 1.78	1.75 1.79	1.74 1.80	1.73 1.81	1.72 1.82	1.71 1.83	1.70 1.84

Учебное издание

РадишевскаяТатьяна Александровна
Шваб Любовь Валентиновна
Щавелева Марина Викторовна
Мещеряков Юрий Владимирович

Математическая статистика
Описательная статистика
Выполнение многомерного регрессионного анализа
(для аспирантов и соискателей ученой степени)

Учебно-методическое пособие

Ответственный за выпуск Ю.В. Мещеряков

Подписано в печать 09. 06. 2016. Формат 60x84/16. Бумага «Discovery».
Печать ризография. Гарнитура «Times New Roman».
Печ. л. 2,09. Уч.- изд. л. 1,62. Тираж 100 экз. Заказ 155
Издатель и полиграфическое исполнение –
Белорусская медицинская академия последипломного образования.
Свидетельство о государственной регистрации издателя, изготовителя,
распространителя печатных изданий № 1/136 от 08.01.2014.
220013, г. Минск, ул. П. Бровки, 3.