

Получена: 3 марта 2017 / Принята: 15 марта 2017 / Опубликовано online: 30 апреля 2017

УДК 614.2 + 303.4

ОДНОФАКТОРНЫЙ ЛИНЕЙНЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ С ИСПОЛЬЗОВАНИЕМ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ STATISTICA И SPSS

Андрей М. Гржибовский¹⁻⁴, <http://orcid.org/0000-0002-5464-0498>

Сергей В. Иванов⁵, <http://orcid.org/0000-0003-0254-3941>

Мария А. Горбатова², <http://orcid.org/0000-0002-6363-9595>

¹ Национальный Институт Общественного Здравоохранения, г. Осло, Норвегия;

² Северный Государственный Медицинский Университет, г. Архангельск, Россия;

³ Международный Казахско-Турецкий Университет им. Х.А. Ясави, г. Туркестан, Казахстан;

⁴ Северо-Восточный Федеральный Университет, г. Якутск, Россия;

⁵ "Первый Санкт-Петербургский государственный медицинский университет им. акад. И.П. Павлова", г. Санкт-Петербург, Россия.

Резюме

В настоящей работе представлены общие сведения о выполнении однофакторного линейного регрессионного анализа данных с использованием программного обеспечения Statistica 10 и SPSS 20 и интерпретация полученных результатов статистической обработки данных. Настоящая статья призвана дать общие сведения об однофакторном линейном регрессионном анализе, и не заменяет прочтения специализированной литературы по статистике и клинической эпидемиологии.

Ключевые слова: Statistica, SPSS, линейная регрессия.

Abstract

UNIVARIATE REGRESSION ANALYSIS USING STATISTICA AND SPSS SOFTWARE

Andrej M. Grjibovski¹⁻⁴, <http://orcid.org/0000-0002-5464-0498>

Sergej V. Ivanov⁵, <http://orcid.org/0000-0003-0254-3941>

Maria A. Gorbatova², <http://orcid.org/0000-0002-6363-9595>

¹ Norwegian Institute of Public Health, Oslo, Norway;

² Northern State Medical University, Arkhangelsk, Russia;

³ International Kazakh-Turkish University, Turkestan, Kazakhstan;

⁴ North-Eastern Federal University, Yakutsk, Russia;

⁵ Pavlov First Saint Petersburg State Medical University, St. Petersburg, Russia.

In this paper we present basic concepts of simple linear regression analysis using Statistica and SPSS software. Special emphasis is given to interpretation of the outputs provided by software packages. The article provides general knowledge and practical advices regarding the use of simple logistic regression in biomedical studies, but it does not substitute specialized literature on biostatistics.

Keywords: Statistica, SPSS, simple linear regression.

Түйіндеме

STATISTICA И SPSS БАҒДАРЛАМАЛЫҚ ҚАМСЫЗДАНДЫРУДЫ ҚОЛДАНУМЕН БІРФАКТОРЛЫ СЫЗЫҚТЫҚ РЕГРЕССИЯЛЫҚ ТАЛДАУ

Андрей М. Гржибовский ¹⁻⁴, <http://orcid.org/0000-0002-5464-0498>**Сергей В. Иванов** ⁵, <http://orcid.org/0000-0003-0254-3941>**Мария А. Горбатова** ², <http://orcid.org/0000-0002-6363-9595>¹ Қоғамдық Денсаулық сақтау Ұлттық Институты, Осло қ., Норвегия;² Солтүстік Мемлекеттік Медициналық Университеті, Архангельск қ., Ресей;³ Х.А. Ясави ат. Халықаралық Қазақ – Түрік Университеті, Туркестан, Қазақстан;⁴ Солтүстік - Шығыс Федералдық Университеті, Якутск қ., Ресей;⁵ Академик И.П. Павлов атынд. бірінші Санкт-Петербург мемлекеттік медициналық университеті, Санкт-Петербург қ., Ресей.

Осы жұмыста Statistica 10 және SPSS 20 бағдарламалық қамсыздандыруды қолданумен бірфакторлы сызықтық регрессиялық талдауды орындау және мәліметтерді статистикалық өңдеуден алынған нәтижелер интерпретация туралы жалпы мәліметтер берілген. Осы мақала бірфакторлы сызықтық регрессиялық талдау туралы жалпы мәліметтер беруге талап етілген және статистика және клиникалық эпидемиология бойынша мамандандырылған әдебиетті оқуды ауыстырмайды.

Негізгі сөздер: Statistica, SPSS, сызықтық регрессия.

Библиографическая ссылка:

Гржибовский А.М., Иванов С.В., Горбатова М.А. Однофакторный линейный регрессионный анализ с использованием программного обеспечения Statistica и SPSS // Наука и Здравоохранение. 2017. №2. С. 5-33.

Grijbovski A.M., Ivanov S.V., Gorbatoва M.A. Univariate regression analysis using Statistica and SPSS software. *Nauka i Zdravookhranenie* [Science & Healthcare]. 2017, 2, pp. 5-33.

Гржибовский А.М., Иванов С.В., Горбатова М.А. Statistica және SPSS бағдарламалық қамсыздандыруды қолданумен бірфакторлы сызықтық регрессиялық талдау // Ғылым және Денсаулық сақтау. 2017. №2. Б. 5-33.

Настоящая статья продолжает серию публикаций, посвященных статистическому анализу данных биомедицинских исследований [10, 15, 16, 17, 18, 19]. Высокое качество статистического анализа является обязательным условием востребованности результатов исследований международным научным сообществом [30, 1], поэтому задачами настоящей серии статей является формирование у начинающего исследователя базисных представлений о статистической обработке данных, приобретение практического опыта работы с современными статистическими пакетами программ и предупреждение типичных ошибок, возникающих в процессе анализа результатов исследования.

Авторы настоятельно рекомендуют читателю еще на этапе планирования научно-исследовательской работы ознакомиться с литературой по эпидемиологии [35, 26, 37], а также с практическими аспектами организации и анализа результатов различных типов научных исследований в здравоохранении (одномоментных, когортных, экологических, экспериментальных исследований и «случай-контроль»), которые представлены в статьях, опубликованных в 2015 году в журнале «Наука и Здравоохранение» [11, 12, 14, 20, 21].

Настоящая статья посвящена однофакторному линейному регрессионному анализу с использованием программного обеспечения Statistica 10 и SPSS 20. Перед изучением статьи авторы настоятельно рекомендуют

читателю ознакомиться с другим методом оценки связи между количественными переменными – корреляционным анализом, который представлен в предыдущей статье серии [13].

При проведении научных исследований в области медицины и общественного здравоохранения, направленных на оценку воздействия различных факторов, на изучаемую систему, одним из конечных результатов анализа является создание математической модели изучаемого явления или процесса. Математические модели позволяют решать широкий спектр задач, стоящих перед исследователем, к числу которых относятся [34]:

- Оценка характера изменения выходных параметров изучаемой системы при изменении входных параметров.

- Оценка степени влияния действующих на систему факторов на ее выходные параметры.

- Прогнозирование уровней выходных параметров системы при заданных значениях действующих факторов.

- Поиск оптимальных уровней факторов для получения требуемых значений выходных параметров.

- Оценка информативности параметров при заданной совокупности действующих на систему факторов.

Главная задача построения модели процесса или явления – помочь исследователю получить дополнительную, неочевидную информацию об изучаемой системе, то есть понять, какие скрытые от непосредственного наблюдения процессы в ней происходят, и понять закономерности, которые напрямую не следует из экспериментальных данных. Соответственно, для того, чтобы выполнить данную задачу, необходимо не только построить модель и получить описывающие ее количественные показатели, но и перевести математические результаты моделирования в понятные, имеющие практическую ценность выводы, понятные специалисту в предметной области и дающие ему новое понимание происходящих в системе процессов [5].

Основные статистические методы моделирования представлены в таблице 1 [2, 6, 25, 27, 29, 36, 38].

Таблица 1.

Основные методы статистического моделирования.

Метод	Переменная отклика	Факторный признак	Примеры
1	2	3	4
Однофакторный регрессионный анализ	Количественная переменная	Один фактор, описываемый количественной переменной	Изучение связи между гестационным возрастом и массой тела новорожденных
Многомерная линейная регрессия	Количественная переменная	2 и более факторов, описываемых количественными переменными	Изучение влияния концентрации оксида углерода и диоксида азота на заболеваемость населения болезнями дыхательной системы с коррекцией на конфаундеры при экологическом дизайне исследования
Однофакторный дисперсионный анализ	Количественная переменная	Один фактор, описываемый качественной переменной	Изучение влияния степени ожирения на уровень систолического артериального давления пациентов
Многофакторный дисперсионный анализ	Количественная переменная	2 и более факторов, описываемых качественными переменными	Изучение влияния социальной группы, уровня образования и наличия сопутствующих заболеваний на показатель качества жизни

Продолжение таблицы 1.

1	2	3	4
Ковариационный анализ	Количественная переменная	2 и более факторов, описываемых и качественными, и количественными переменными	Изучение влияния пола, возраста и индекса массы тела пациентов на уровень холестерина в крови
Логлинейный анализ	Качественная переменная	2 и более факторов, описываемых качественными переменными	Изучение влияния пола, степени тяжести заболевания и наличия сопутствующих заболеваний на выживаемость пациентов в течение определенного периода
Бинарная логистическая регрессия	Дихотомическая (бинарная) качественная переменная	2 и более факторов, описываемых и качественными, и количественными переменными	Изучение влияния пола, возраста пациентов и степени тяжести заболевания на факт возникновения определенного осложнения заболевания
Мультиномиальная логистическая регрессия	Качественная переменная	2 и более факторов, описываемых качественными и ординальными переменными	Изучение влияния пола, возрастной категории пациентов и степени приверженности к терапии на тип течения заболевания (в случае 3-х и более взаимоисключающих типов течения заболевания)
Порядковая регрессия	Порядковая (ранговая) переменная	2 и более факторов, описываемых качественными и ординальными переменными	Изучение влияния пола, возрастной группы и степени ожирения пациентов на степень артериальной гипертензии

Однофакторный линейный регрессионный анализ является относительно простым методом, наиболее предпочтительным для получения базовых представлений об основных принципах статистического моделирования. В последующих статьях серии будут представлены методы моделирования, наиболее часто встречающиеся в биомедицинских исследованиях – множественная (многомерная, многофакторная) линейная регрессия и логистическая регрессия.

С остальными методами моделирования читатель может ознакомиться самостоятельно в процессе прочтения специализированной литературы по статистическому анализу [5, 2, 6, 25, 27, 29, 36, 38].

Для получения наглядного представления об однофакторном линейном регрессионном анализе рассмотрим пример линейной связи между условным независимым параметром X

и зависимым параметром Y , представленным для 10-ти наблюдений (рисунок 1).

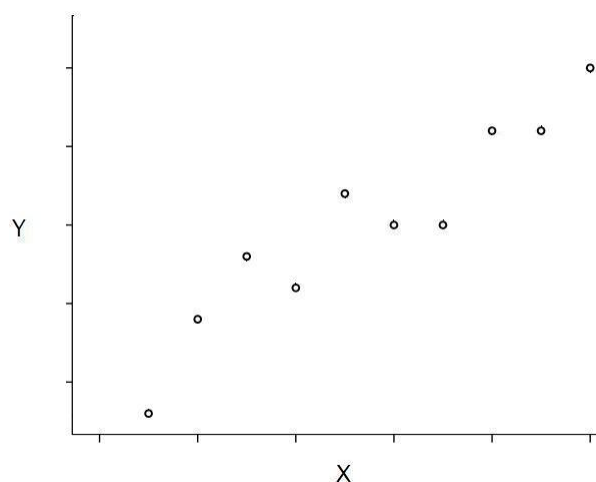


Рис. 1. Скаттерограмма зависимости между параметрами X и Y .

Линейный характер связи в данном случае очевиден, что позволяет использовать однофакторный регрессионный анализ для моделирования зависимости параметра Y от параметра X . Так как связь носит линейный характер, она может быть описана уравнением линейной функции вида $Y = a \times X + b$.

Но само по себе уравнение данного вида подходит только для случая, когда все точки располагаются на прямой, описываемой данным уравнением, то есть для случая функциональной, а не корреляционной связи между переменными.

На рисунке 1 видно, что точки не располагаются на одной линии, но зависимость между переменными наиболее точно может быть отражена с помощью прямой, расположенной таким образом, чтобы все точки располагались максимальной близко к данной прямой. Такая прямая представлена на рисунке 2, и значение параметра Y в зависимости от параметра X будет описано уравнением $Y = b_1 \times X + b_0 + \varepsilon$, где b_1 – коэффициент модели, b_0 – константа модели, а ε – случайная ошибка модели. Если бы между зависимой переменной Y и независимой переменной X не было бы никакой связи, то регрессионная прямая проходила бы параллельно оси абсцисс через среднее значение переменной Y .

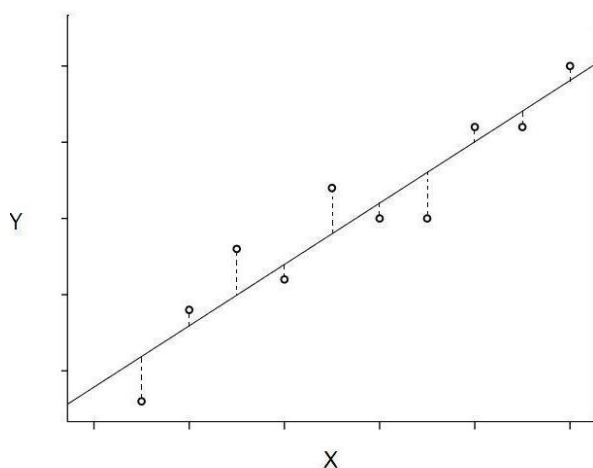


Рис. 2. Построение линии регрессии для параметров X и Y .

Следует отметить, что связь между количественными переменными также можно оценить с помощью корреляционного анализа, используя коэффициент корреляции Пирсона

или Спирмена [13, 7, 34]. При этом однофакторный линейный регрессионный анализ в определенном смысле можно рассматривать как углубленный анализ взаимосвязи переменных, позволяющий прогнозировать значение одной (зависимой) переменной на основании другой (независимой) переменной.

На рисунке 2 представлена линия, описываемая данным уравнением регрессии $Y = b_1 \times X + b_0 + \varepsilon$ (пунктирными линиями указаны «остатки» – разница между фактическим и представленным на линии значением переменной Y).

Степень выраженности разброса наблюдений вокруг линии регрессии определяется показателем, который называется суммой квадратов остатков (SS_r).

SS_r отражает меру неточности модели, или, другими словами, меру изменчивости (вариации) зависимой переменной Y , которую невозможно объяснить независимой переменной X . Линия регрессии проводится таким образом, чтобы значение SS_r оказалось наименьшим. Почему же для оценки точности модели используется не просто сумма значений остатков, а именно сумма квадратов остатков?

Дело в том, что остатки, расположенные выше линии регрессии, будут иметь положительный знак, а расположенные ниже – отрицательный, и сумма значений остатков так или иначе будет близка к нулю и не даст никакой информации о том, насколько далеко расположены фактические значения зависимой переменной от прогнозируемых (предсказанных) значений, расположенных регрессионной прямой. В данном случае именно сумма квадратов остатков позволяет оценить совокупную степень удаленности фактических значений от предсказанных, так как ликвидирует влияние знаков остатков.

Соответственно, качество модели определяется тем, насколько точно она способна предсказать значение зависимой переменной Y на основании независимой переменной X . Вариацию зависимой переменной Y , объясняемую построенной линейной регрессионной моделью, можно представить в виде формулы: $SS_m = SS_t - SS_r$, где SS_t – общая сумма квадратов для всех

значений зависимой переменной Y (мера общей вариации переменной Y), а SS_r – представленная выше сумма квадратов остатков [8].

Одним из главных показателей качества модели является доля вариации зависимой переменной, которую способна объяснить модель: чем ближе она к 100%, тем выше качество модели. Доля общей вариации, которую способна объяснить регрессионная модель, выражается в виде коэффициента детерминации (R^2), который рассчитывается по формуле $R^2 = SS_m / SS_t$.

На практике ни одна из моделей не имеет 100% точности, так как такая степень точности предполагает наличие функциональной связи между переменными, что не наблюдается в изучаемых в медицине и общественном здравоохранении процессах. Следует отметить, что оценка степени неточности проводится для всех моделей, построенных с использованием различных методов математического моделирования (таблица 1), но в каждом случае используется определенный способ оценки неточности модели. К тому же все методы моделирования чувствительны к объему выборки наблюдений – чем больше объем выборки, тем более высокое качество модели следует ожидать.

Построение регрессионной линии методом наименьших квадратов вручную и ручной подбор соответствующих коэффициента и константы для уравнения $y = b_1 \times x + b_0$ крайне затруднителен, поэтому однофакторный регрессионный анализ, как и остальные методы статистического моделирования, проводится с использованием специального статистического программного обеспечения, о котором будет сказано ниже. Именно простота получения уравнений регрессии с помощью современных статистических программ делает регрессионный анализ доступными для большого числа исследователей, но данный факт часто играет негативную роль, так как простота построения модели часто несопоставима со сложностью практической интерпретации построенной модели, и поэтому исследователи часто допускают серьезные ошибки в именно на этапе интерпретации моделей [5, 22].

Как и любой другой статистический метод, однофакторный линейный регрессионный анализ имеет определенные условия применения [8]:

- Выборка из популяции должна быть репрезентативной.
- Зависимая переменная должна быть количественной.
- Независимая переменная должна быть количественной.
- Зависимость между переменными должна носить линейный характер (проверяется графически путем построения скаттерограммы).
- Наблюдения (и остатки) должны быть независимы друг от друга (проверяется с помощью критерия Durbin-Watson).
- Остатки должны иметь нормальное распределение (проверяется графически путем построения гистограмм и квантильных диаграмм, а также с помощью критериев Колмогорова-Смирнова или Шапиро-Уилка). О проверке нормальности распределения подробно сказано в первой статье настоящего цикла публикаций [15].
- Остатки должны иметь одинаковый разброс на всем протяжении предсказанных значений (или независимой переменной). Данное условие проверяется путем построения скаттерограммы (в идеальной ситуации график должен представлять собой бессистемный разброс точек).

Первые четыре условия применения метода не требуют комментариев по причине очевидности, а остальные условия будут подробно описаны ниже, в процессе пошагового описания алгоритма анализа.

Следует обратить внимание исследователей на первое условие применения однофакторного линейного регрессионного анализа, так как модель, удовлетворяющая всем остальным условиям и имеющая высокую внутреннюю валидность, может оказаться совершенно бесполезной в отношении применения на уровне генеральной совокупности, если выборка наблюдений не является репрезентативной.

Для того, чтобы читатель приобрел практические навыки выполнения однофакторного линейного регрессионного анализа, будет рассмотрен фрагмент данных,

которые были собраны в ходе исследования, направленного на изучение метаболического синдрома и его детерминант в условиях неблагополучной социально-экологической ситуации в Южном Казахстане [23, 24, 28, 32].

В ходе данного исследования были получены значения индекса массы тела (ИМТ) и окружности талии 277 пациентов.

Однофакторный линейный регрессионный анализ будет проведен с использованием программного обеспечения Statistica 10 [31, 3] и SPSS 20 [4], демонстрационные версии которого можно загрузить с официальных сайтов разработчиков (www.statsoft.com и www.ibm.com соответственно).

Представленные ниже алгоритмы действий являются не более чем инструментом анализа данных, в то время как корректная интерпретация полученных результатов требует наличия базисных знаний в области биомедицинской статистики, которые могут быть получены только путем изучения специализированной литературы [5, 2, 6, 25, 27, 29, 36, 38].

Однофакторный линейный регрессионный анализ с использованием программы Statistica 10

Для начала работы необходимо открыть файл 8_Regression_STAT.sta, который потребуется загрузить с сайта журнала «Наука и Здравоохранение». В данном файле представлены 2 непрерывные количественные

переменные: ИМТ (переменная «BMI») и окружность талии (переменная «Waist_circum»).

В результате статистического анализа данных будет построена линейная регрессионная модель зависимости ИМТ от окружности талии пациентов с метаболическим синдромом. Таким образом, переменная «BMI» выступает в роли зависимой переменной, а переменная «Waist_circum» является независимой переменной (переменной-предиктором).

На начальном этапе обработки данных требуется построить скаттерограмму, чтобы визуально определить, носит ли связь между переменными линейный характер.

Для этого войдем в меню «Graphs» в верхней части экрана и выберем раздел «Scatterplots...». В появившемся окне «2D Scatterplots» (рисунок 3) нажмем на кнопку «Variables» и выберем переменные, значения которых будут отложены по осям абсцисс и ординат, как это показано на рисунке 4. Выбор подтвердим нажатием на кнопку «OK» и в открывшемся окне «2D Scatterplots» снова нажмем на кнопку «OK».

В результате наших действий программа сформирует скаттерограмму зависимости между переменными «BMI» и «Waist_circum» (рисунок 5). Видно, что скаттерограмма отражает линейную зависимость и в достаточной мере гомоскедастична: разброс значений одной переменной практически не зависит от разброса другой переменной.

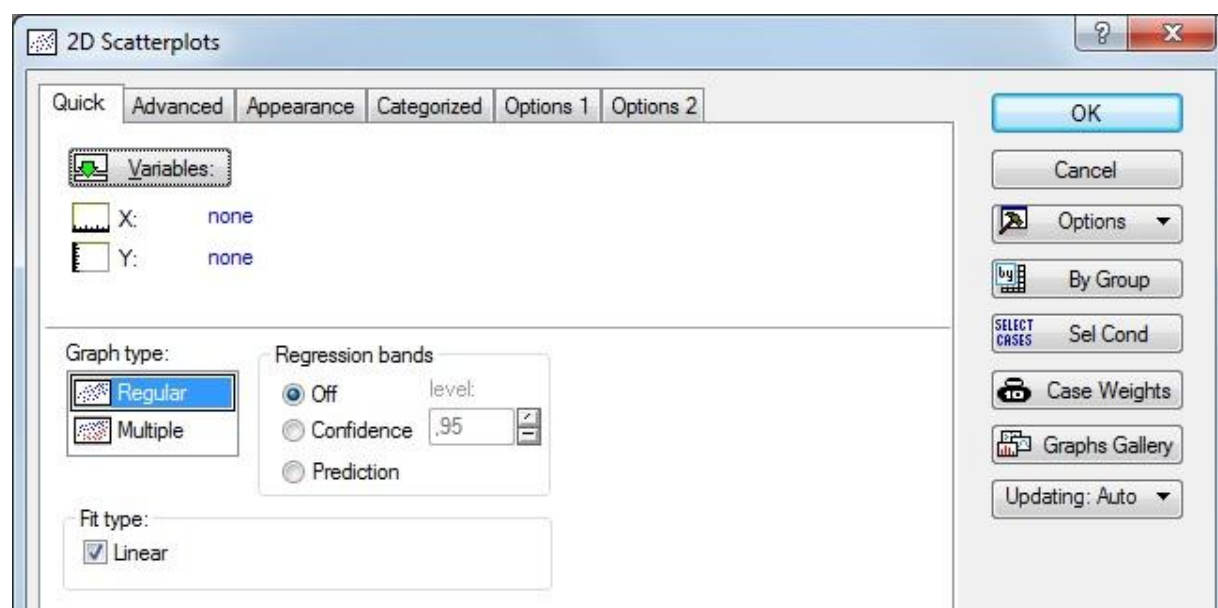


Рис. 3. Окно «2D Scatterplots» программы Statistica 10.

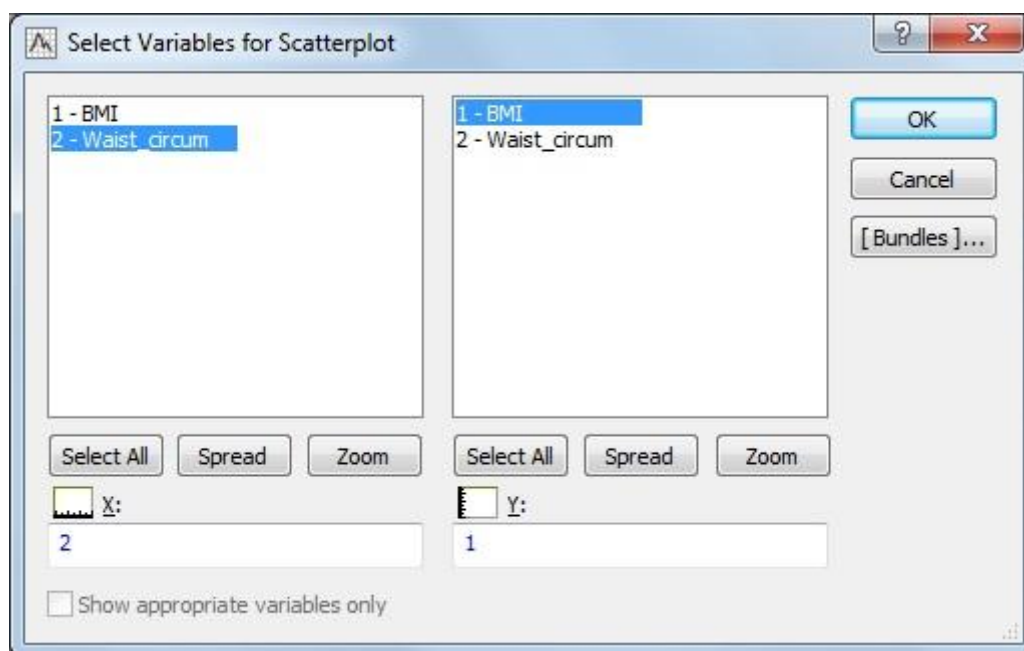


Рис. 4. Окно «Select Variables for Scatterplot» программы Statistica 10 (выбор переменных «Waist_circum» и «BMI»).

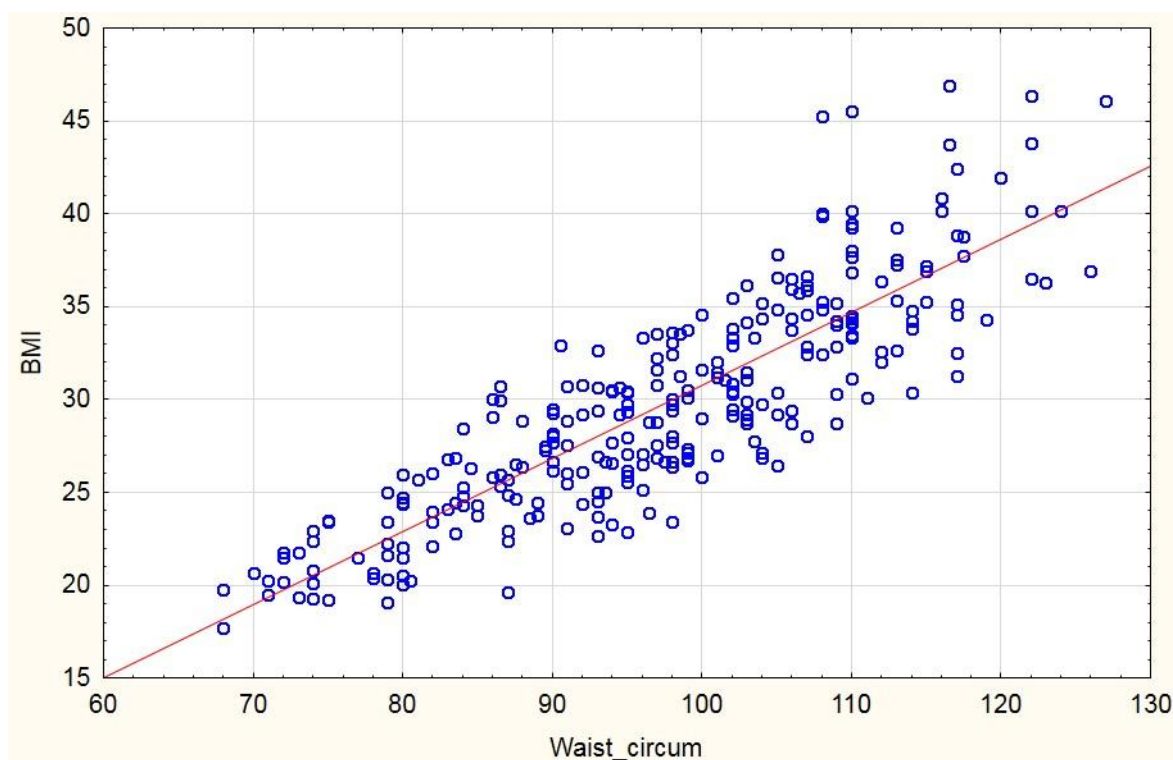


Рис. 5. Скаттерограмма зависимости переменной «BMI» от переменной «Waist_circum» (программа Statistica 10).

На следующем этапе обработки данных для проведения однофакторного линейного регрессионного анализа выбираем меню

«Statistics» (в верхней части рабочего пространства программы) и входим в раздел «Multiple Regression» (рисунок 6).

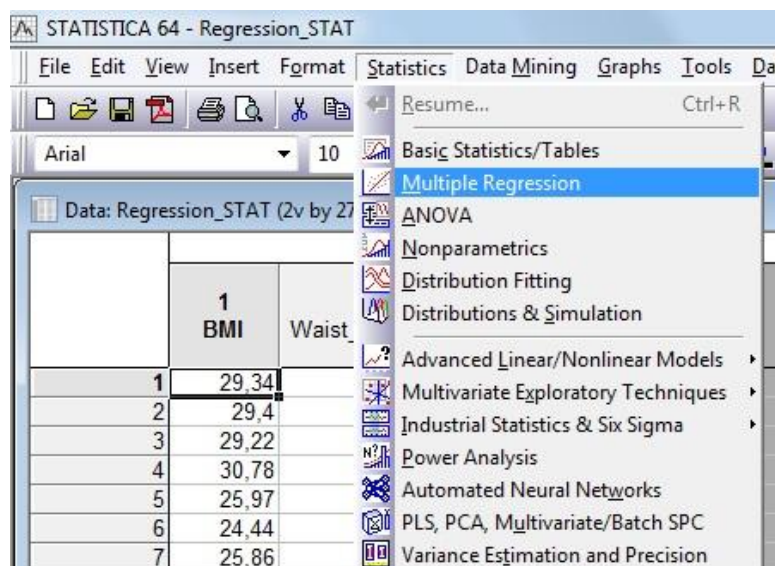


Рис. 6. Выбор раздела «Multiple Regression» меню «Statistics» программы Statistica 10.

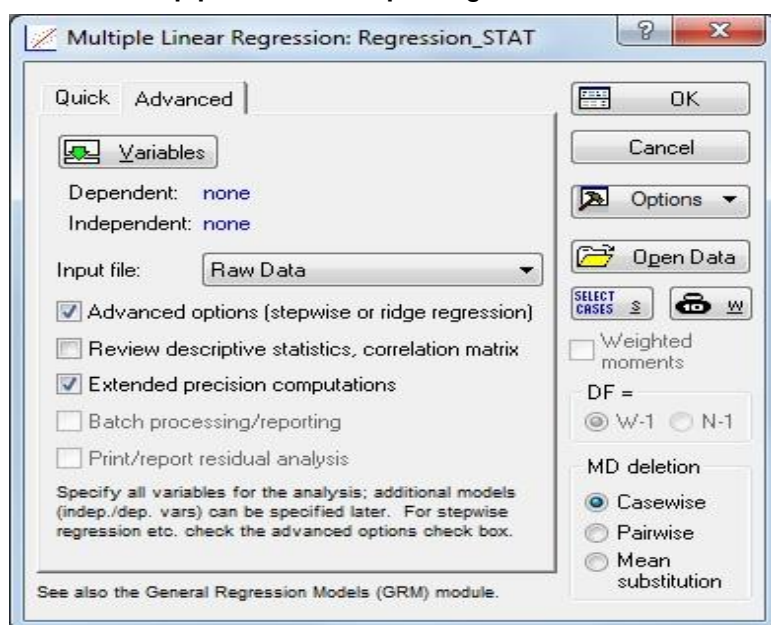


Рис. 7. Окно «Multiple Linear Regression» программы Statistica 10.

В открывшемся окне сразу переключаемся на вкладку «Advanced» и отмечаем галочками позиции «Advanced options (stepwise or ridge regression)» и «Extended precision computations», которые в дальнейшем откроют возможность углубленного анализа. Далее нажимаем на кнопку «Variables», чтобы выбрать переменные для анализа (рисунок 7).

Программа откроет окно, в котором выберем переменную «BMI» как зависимую (поле «Dependent var. (or list for batch)»), а переменную «Waist_circum» – как независимую (поле «Independent variable list»), подтвердив выбор нажатием на кнопку «OK» (рисунок 8).

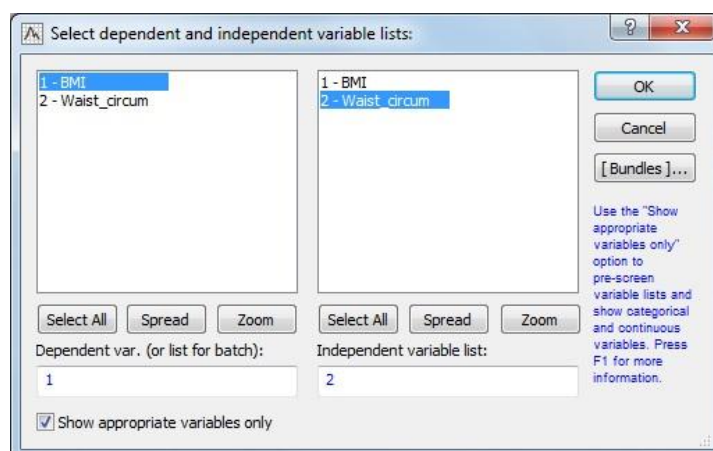


Рис. 8. Окно «Select dependent and independent variable lists» программы Statistica 10 (выбор переменных «BMI» и «Waist_circum»).

Программа вернется к окну «Multiple Linear Regression» (рисунок 7), в котором нажмем на кнопку «OK», после чего откроется окно «Model Definition», в котором снова нажмем на кнопку «OK» (рисунок 9). Данное окно позволяет задавать параметры модели, но, так как мы рассматриваем самый простой вариант линейной регрессионной модели – с одним независимым фактором, внесение изменений в изначальные настройки модели не требуется.

Далее программа откроет окно «Multiple regression results», с помощью которого мы будем пошагово выводить результаты моделирования и проверять степень соответствия построенной модели фактическим данным.

В окне «Multiple regression results» переключимся на вкладку «Advanced» (рисунок 10) и нажмем на кнопку «Summary»:

Regression results» для вывода основных показателей модели (рисунок 11).

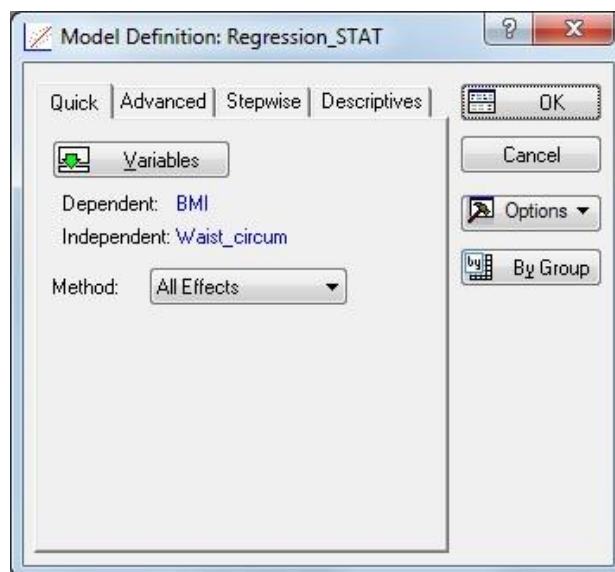


Рис. 9. Окно «Model Definition» программы Statistica 10.

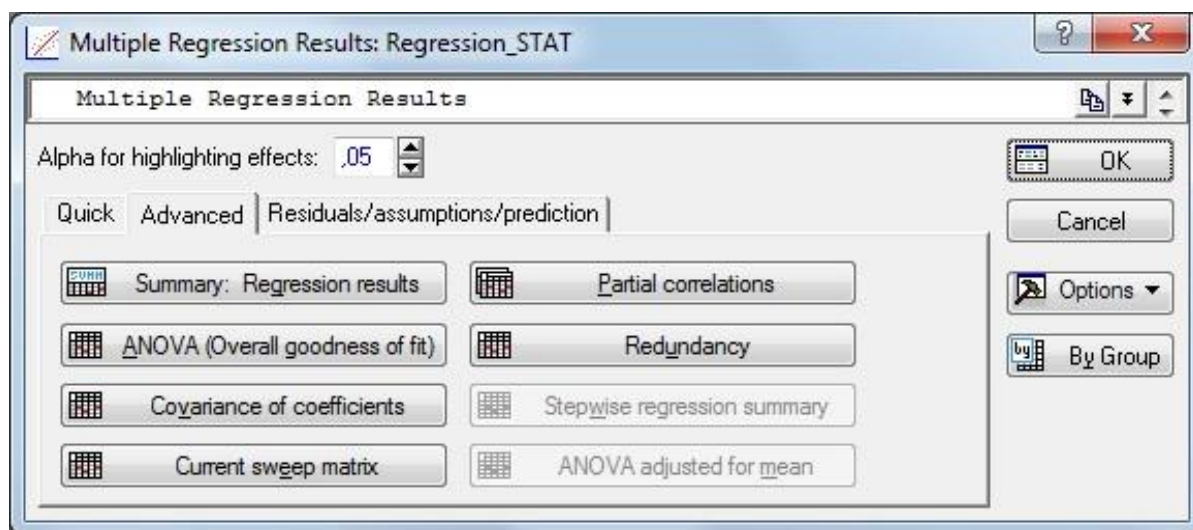


Рис. 10. Вкладка «Advanced» окна «Multiple regression results» программы Statistica 10.

Workbook1* - Regression Summary for Dependent Variable: BMI (Regression_STAT)

Regression Summary for Dependent Variable: BMI (Regression_STAT)						
R= ,85705243 R²= ,73453887 Adjusted R²= ,73357355 F(1,275)=760,93 p<0,0000 Std.Error of estimate: 3,0387						
	b*	Std.Err. of b*	b	Std.Err. of b	t(275)	p-value
Intercept			-8,60319	1,401966	-6,13652	0,000000
Waist_circum	0,857052	0,031069	0,39360	0,014269	27,58502	0,000000

Рис. 11. Общие сведения о линейной регрессионной модели, описывающей связь между зависимой переменной «BMI» и независимой переменной «Waist_circum».

На рисунке 11 представлены следующая информация о модели:

1. Статистическая значимость модели, проверяемая с помощью критерия Фишера (F): $F_{1, 275} = 760,93$, $p < 0,0001$. Так как $p < 0,0001$ (программа указывает « $p < 0,0000$ »), то нулевую гипотезу об отсутствии взаимосвязи между переменными «BMI» и «Waist_circum» можно отвергнуть, то есть статистически подтверждено наличие связи между изучаемыми переменными.

2. Коэффициент детерминации модели: $R^2 = 0,734$.

3. Значения константы (b_0) и коэффициента регрессии (b_1) уравнения линейной регрессии $y = b_1 \times x + b_0$: $b_0 = -8,603$ (графа «b», первая строка), $b_1 = 0,394$ (графа «b», вторая строка). Программа также с помощью коэффициента Стьюдента проверяет нулевую гипотезу о равенстве нулю значений коэффициента и константы. В данном случае значения коэффициента Стьюдента (графа «t(275)») позволяют отвергнуть нулевую гипотезу как в отношении константы, так и в отношении коэффициента регрессии (представленные в графе «p-value» уровни статистической значимости меньше 0,05 как для константы, так и для коэффициента регрессии).

Таким образом, модель статистически значима, зависимость значения переменной «BMI» от значения переменной «Waist_circum» описывается уравнением $Y_{BMI} = -8,603 + 0,394 \times X_{Waist_circum}$, и доля вариации зависимой переменной «BMI», которую способна объяснить модель на основании независимой переменной «Waist_circum», составляет 73,4%.

Значение коэффициента регрессии (b_1) показывает, насколько увеличится значение зависимой переменной при увеличении независимой переменной на единицу. Для данного примера ИМТ увеличивается на 0,394 кг/м² при увеличении окружности талии на 1 см. Существует также стандартизованный коэффициент регрессии, который показывает, на сколько стандартных отклонений увеличится значение зависимой переменной при увеличении независимой переменной на одно стандартное отклонение (но на практике стандартизованные коэффициенты используются нечасто). Константа (b_0)

показывает, в какой точке регрессионная прямая пересечет ось ординат, то есть показывает значение зависимой переменной, если значение независимой переменной равно нулю. Для приведенного примера константа не имеет практического смысла, так как ИМТ не может равняться -8,603 кг/м², да само по себе прогнозирование ИМТ для ситуаций, когда окружность талии стремится к нулю, является абсурдным. Рассмотренная ситуация наглядно показывает, что прогнозировать значения зависимой переменной следует только для того диапазона независимой переменной, на основании которого было построено уравнение регрессии. Так, в данном примере прогнозировать ИМТ пациентов с помощью вышеприведенного уравнения целесообразно только при значениях окружности талии от 68 до 127 см, несмотря на то, что регрессионную прямую можно провести далеко за пределы имеющихся данных в оба направления. Следует отметить, что подобное экстраполирование является довольно частой ошибкой исследователей, когда проводится прогнозирование значения зависимой переменной по значениям независимой переменной, которые не входили в исследование. Хотелось бы предостеречь начинающих и опытных исследователей от таких прогнозов, так как построенная линейная модель совершенно не гарантирует сохранения линейного характера зависимости на всем протяжении регрессионной прямой.

Возвращаясь к окнам программы Statistica 10, следует отметить, что в дальнейшем для просмотра результатов анализа, появляющихся при выполнении пошагового алгоритма, переключаться между окнами вывода данных можно с помощью дерева каталогов в левой части рабочего окна (рисунок 11), а возвращаться к процессу анализа следует путем нажатия на вкладку «Multiple Linear Regression» в нижнем левом углу рабочего поля программы.

Итак, нажмем на вкладку «Multiple Linear Regression» для возврата к процессу вывода результатов анализа. Программа откроет окно «Multiple regression results» (рисунок 10), в котором выберем вкладку «Residuals/assumptions/prediction» (рисунок 12) и нажмем на кнопку «Perform residual analysis».

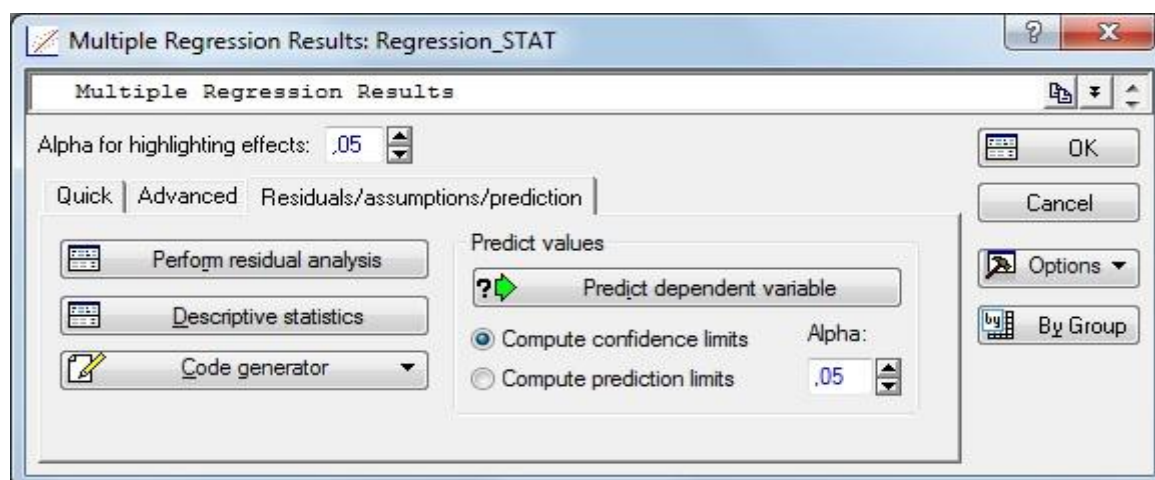


Рис. 12. Вкладка «Residuals/assumptions/prediction» окна «Multiple regression results» программы Statistica 10.

Программа откроет следующее окно на вкладку «Advanced» и нажмем на кнопку «Residual Analysis», в котором переключимся «Durbin-Watson statistic» (рисунок 13).

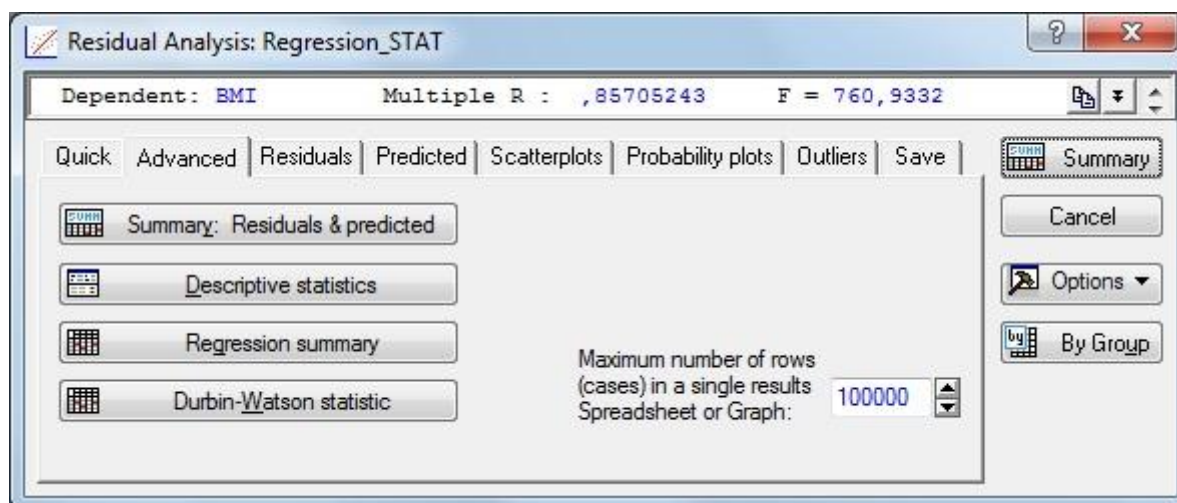


Рис. 13. Вкладка «Advanced» окна «Residual Analysis» программы Statistica 10.

Результаты расчета критерия Durbin-Watson представлены на рисунке 14. Данный критерий проверяет условие независимости наблюдений друг от друга – обязательное условие применения однофакторного линейного регрессионного анализа. Допустимые значения для данного критерия – от 1 до 3. Если данный критерий имеет

значение менее 1 или более 3, это означает, что условие независимости наблюдений не соблюдается, и прогнозирование зависимой переменной с помощью однофакторного линейного регрессионного анализа не может считаться корректным [8]. В нашем случае данное условие соблюдено, так как значение критерия Durbin-Watson близко к 2.

Workbook2* - Durbin-Watson d (Regression_STAT)			
Workbook2*			
Multiple Regression			
Regression result			
Summary Statistics			
Regression Statistics			
Regression Residuals			
Durbin-Watson			
Durbin-Watson d (Regression_STAT) and serial correlation of residuals			
	Durbin-Watson d	Serial Corr.	
Estimate	1,934426	0,032720	

Рис. 14. Результаты расчета критерия Durbin-Watson.

Далее снова нажмем на вкладку «Multiple Linear Regression» в нижнем левом углу рабочего поля программы. Программа вернется к окну «Residual Analysis» (рисунок

13), в котором нажмем на вкладку «Residuals», после чего нажмем на кнопку «Histogram of Residuals» для вывода гистограммы распределения остатков модели (рисунок 15).

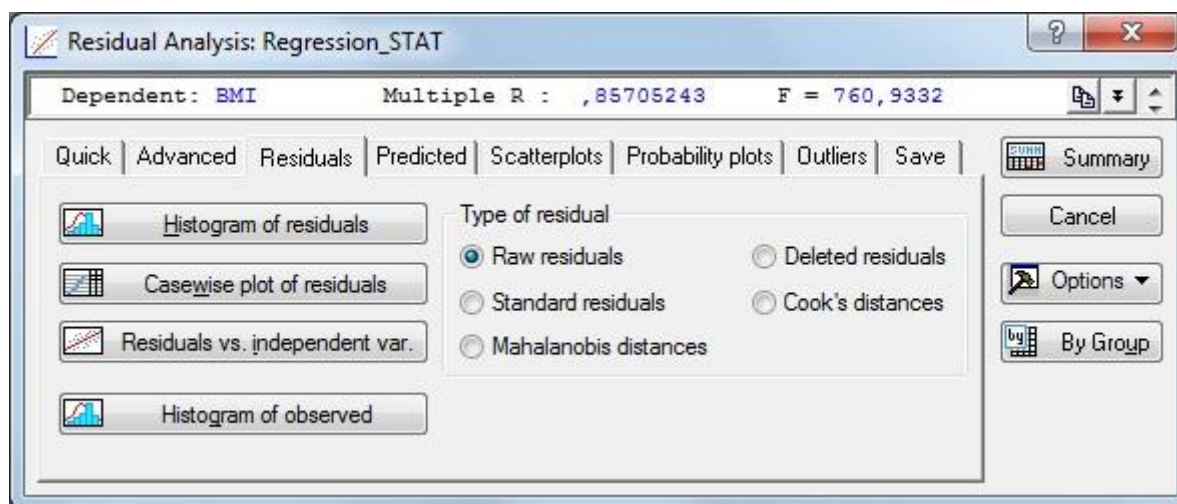


Рис. 15. Вкладка «Residuals» окна «Residual Analysis» программы Statistica 10.

В результате наших действий программа представит гистограмму распределения остатков модели (рисунок 16). Мы видим, что гистограмма имеет колоколообразную форму, близкую к нормальному распределению. Построение гистограмм является только одним из способов проверки распределения, другие способы включают в себя построение

квантильной диаграммы и расчет значений критериев Колмогорова-Смирнова и Шапиро-Уилка [15,9,33]. Итак, на основании гистограммы распределения мы можем сделать вывод о том, что остатки модели имеют распределение, близкое к нормальному, то есть данное условие применения однофакторного линейного регрессионного анализа соблюдено.

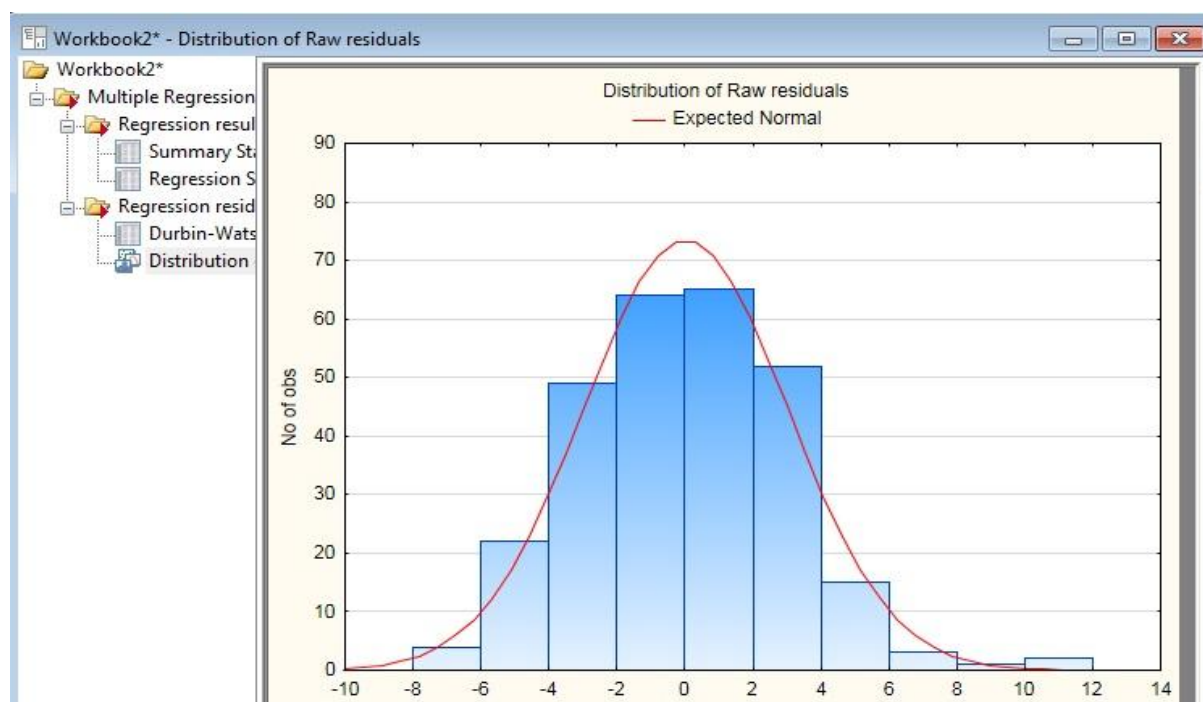


Рис. 16. Гистограмма распределения остатков модели (программа Statistica 10).

Далее вернемся к окну «Residual Analysis» и переключимся на вкладку «Outliers» для анализа «выбросов» – имеющих в выборке

атипичных наблюдений (рисунок 17). По умолчанию программа считает «выбросами» все значения, отклоняющиеся от среднего

более чем на 2 стандартных отклонения («Standard residual ($> 2 * \sigma$)»). Нажатие на кнопку «Casewise plot of outliers» запустит вывод таблицы с перечнем «выбросов» (рисунок 18).

Согласно определению нормального распределения, количество подобных атипичных наблюдений не должно составлять более 5% от объема выборки. В нашем случае количество наблюдений, для которых остатки имеют значения либо менее 2, либо более 2 стандартных отклонений, составило 9 единиц,

то есть $9/277 = 3,2\%$ от общего объема выборки, что укладывается в ожидаемые 5%. Следует отметить, что количество наблюдений с большими отрицательными остатками должно быть приблизительно равно количеству наблюдений с большими положительными остатками, что говорит о симметричности распределения остатков. В нашем случае это условие не соблюдается, но относительно незначительное количество «выбросов» (3,2% от объема выборки) нивелирует данное несоответствие.

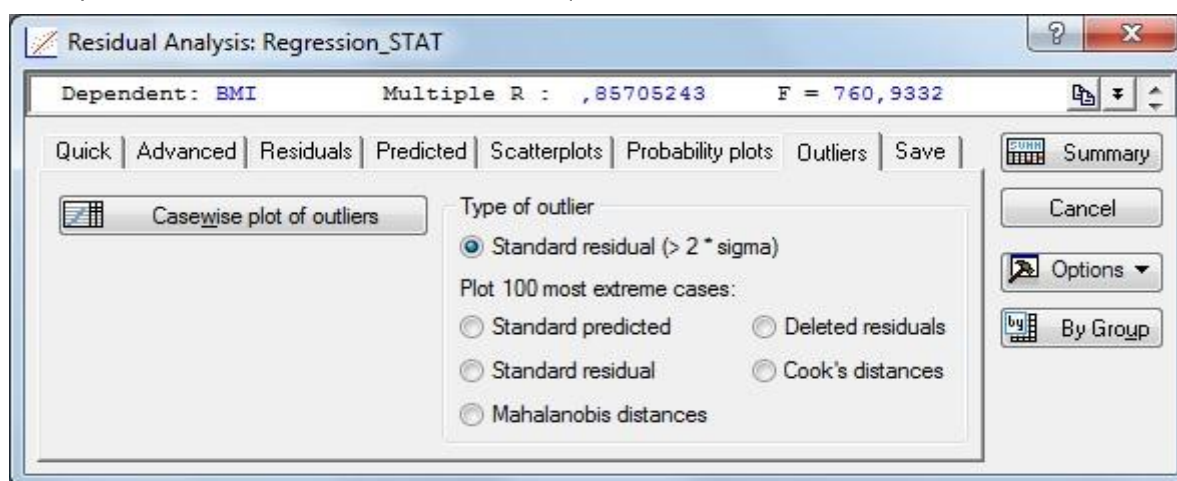


Рис. 17. Вкладка «Outliers» окна «Residual Analysis» программы Statistica 10.

Standard Residuals									
Case	-5.	-4.	-3.	±2.	3.	4.	5.		
14	.	.	.	*	.	.	.	26,42000	32,72469
106	.	.	.		.	*	.	45,27000	33,90548
107	.	.	.		*	.	.	46,87000	37,25107
172	.	.	.		.	*	.	45,53000	34,69268
186	.	.	.	*	.	.	.	31,23000	37,44787
243	.	.	.	*	.	.	.	43,71000	37,25107
247	.	.	.	*	.	.	.	40,01000	33,90548
263	.	.	.	*	.	.	.	46,36000	39,41586
268	.	.	.	*	.	.	.	23,44000	29,96949
Minimum	.	.	.		.	.	.	23,44000	29,96949
Maximum	.	.	.		.	*	.	46,87000	39,41586
Mean	.	.	.	*	.	.	.	38,76000	35,17374
Median	.	.	.	*	.	.	.	43,71000	34,69268

Outliers						
Observed Value	Predicted Value	Residual	Standard Pred. v.	Standard Residual	Std. Err. Pred. Val	
26,42000	32,72469	-6,30469	0,591404	-2,07478	0,212218	
45,27000	33,90548	11,36452	0,825431	3,73991	0,236917	
46,87000	37,25107	9,61893	1,488508	3,16546	0,327813	
45,53000	34,69268	10,83732	0,981449	3,56641	0,256049	
31,23000	37,44787	-6,21787	1,527512	-2,04622	0,333762	
43,71000	37,25107	6,45893	1,488508	2,12554	0,327813	
40,01000	33,90548	6,10452	0,825431	2,00891	0,236917	
46,36000	39,41586	6,94414	1,917557	2,28522	0,395415	
23,44000	29,96949	-6,52949	0,045341	-2,14877	0,182767	
23,44000	29,96949	-6,52949	0,045341	-2,14877	0,182767	
46,87000	39,41586	11,36452	1,917557	3,73991	0,395415	
38,76000	35,17374	3,58626	1,076793	1,18019	0,278852	
43,71000	34,69268	6,45893	0,981449	2,12554	0,256049	

Рис. 18. Результаты анализа «выбросов» в выборке.

В том случае, если в выборке имеется большое количество наблюдений с остатками более 3 стандартных отклонений, это факт может свидетельствовать о большом разбросе данных вокруг регрессионной прямой, то есть о недостаточном соответствии модели имеющимся данным. А в том случае, если большинство из атипичных остатков имеют один и тот же знак, данный факт может говорить о наличии кластера наблюдений, которые не описываются построенной моделью.

Далее вернемся к анализу, и в окне «Residual Analysis» выберем вкладку «Save» и нажмем на кнопку «Save residuals & predicted» для отображения таблицы с прогнозируемыми значениями переменной «BMI», остатками и другими расчетными параметрами по каждому наблюдению выборки (рисунок 19).

В открывшемся окне «Select variables to save with predicted/resid...» ничего не выбираем и нажимаем на кнопку «OK».

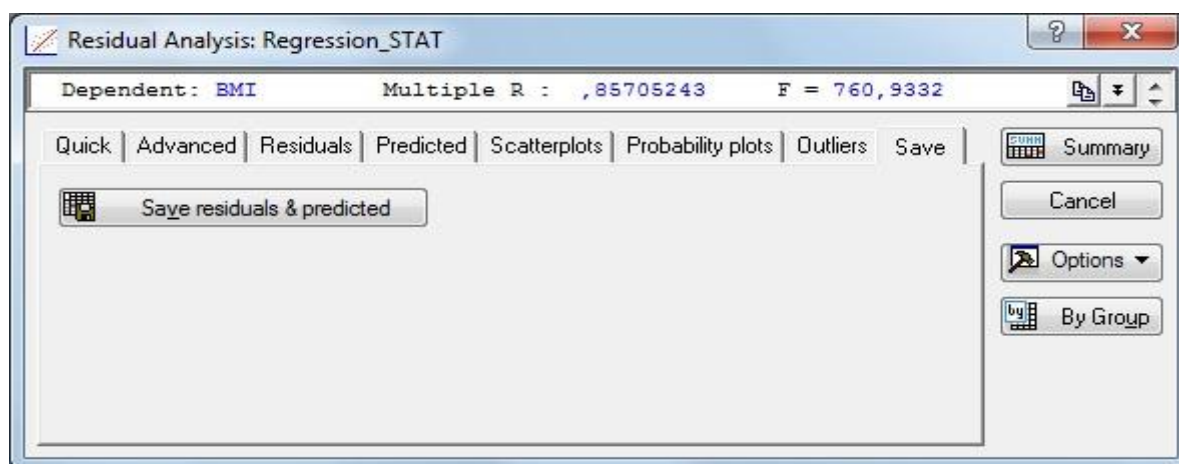


Рис. 19. Вкладка «Save» окна «Residual Analysis» программы Statistica 10.

Программа представит таблицу (рисунок 20), в которой нас, в первую очередь, интересуют прогнозируемые на основании модели значения переменной «BMI» (графа «1 Predicted»), остатки (графа «2 Residuals»),

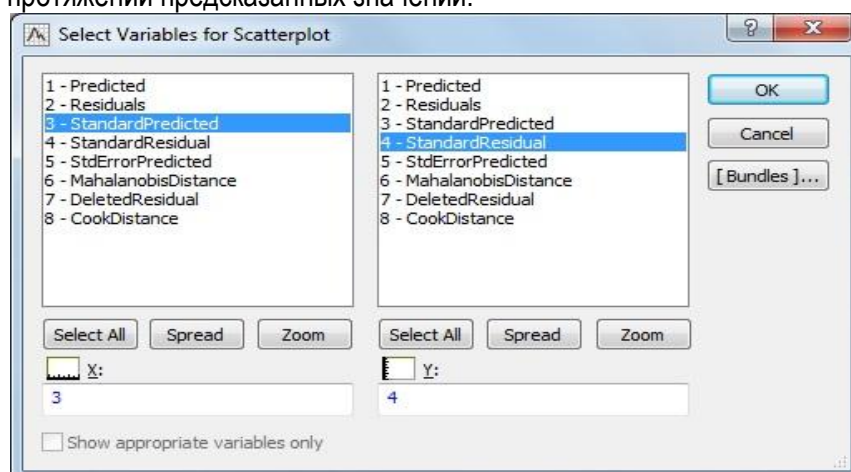
стандартизированные прогнозируемые значения и стандартизированные остатки (графы «3 StandardPredicted» и «4 StandardResidual» соответственно).

Regression_STAT								
	1 Predicted	2 Residuals	3 StandardPredicted	4 StandardResidual	5 StdErrorPredicted	6 MahalanobisDistance	7 DeletedResidual	8 CookDistance
1	28.79	0.55	-0.19	0.18	0.19	0.04	0.55	0.00
2	33.12	-3.72	0.67	-1.22	0.22	0.45	-3.74	0.00
3	32.72	-3.50	0.59	-1.15	0.21	0.35	-3.52	0.00
4	29.58	1.20	-0.03	0.40	0.18	0.00	1.21	0.00
5	22.88	3.09	-1.36	1.02	0.31	1.85	3.12	0.01
6	26.43	-1.99	-0.66	-0.65	0.22	0.43	-2.00	0.00
7	28.79	-2.93	-0.19	-0.96	0.19	0.04	-2.94	0.00
8	34.69	1.69	0.98	1.51	0.26	0.96	1.62	0.01

Рис. 20. Таблица прогнозируемых, на основании модели значений переменной «BMI», остатков, стандартизированных прогнозируемых значений и стандартизированных остатков.

С помощью данной таблицы проверим последнее условие применения однофакторного линейного регрессионного анализа, согласно которому остатки должны иметь одинаковый разброс на всем протяжении предсказанных значений.

Для этого войдем в меню «Graphs» в верхней части экрана и выберем раздел «Scatterplots...» так же, как при построении скаттерограммы для оценки линейного характера связи между переменными.



В появившемся окне «2D Scatterplots» (рисунок 3) нажмем на кнопку «Variables» и выберем переменные, значения которых будут отложены по осям абсцисс и ординат, как показано на рисунке 21. Выбор подтвердим нажатием на кнопку «OK» и в открывшемся окне «2D Scatterplots» снова нажмем на кнопку «OK».

Рис. 21. Окно «Select Variables for Scatterplot» программы Statistica 10 (выбор переменных «StandardPredicted» и «StandardResidual»).

В результате наших действий программа сформирует скаттерограмму зависимости между стандартизированными прогнозируемыми значениями переменной «BMI» и стандартизированными остатками (рисунок 22). Видно, что скаттерограмма представляет собой бессистемный разброс

точек, то есть остатки имеют примерно одинаковый разброс на всем протяжении предсказанных значений. Таким образом, можно сделать вывод о том, что последнее условие применения однофакторного линейного регрессионного анализа соблюдено.

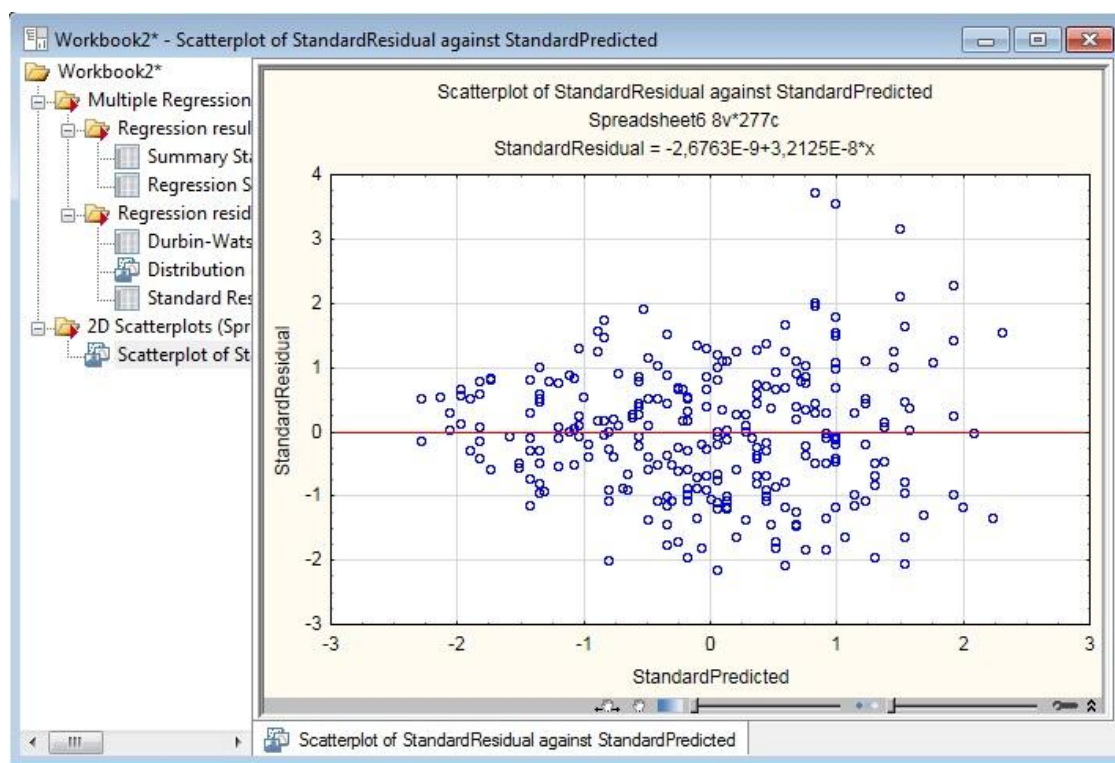


Рис. 22. Скаттерограмма зависимости между стандартизированными прогнозируемыми значениями переменной «BMI» и стандартизированными остатками (программа Statistica 10).

Таким образом, построенная статистическая модель, описывающая зависимость ИМТ от окружности талии пациентов с метаболическим синдромом, имеет достаточную внутреннюю валидность, о чем можно судить на основании значения коэффициента детерминации, равного 0,734, и соблюдении всех условий применения однофакторного линейного регрессионного анализа. Модель также имеет достаточную внешнюю валидность, так как имеющаяся выборка наблюдений является сплошной.

Следует отметить, что регрессионное уравнение модели $Y_{\text{BMI}} = -8,603 + 0,394 \times X_{\text{Waist_circum}}$ позволяет прогнозировать только среднее значение ИМТ: например, среднее значение ИМТ пациентов с метаболическим синдромом, имеющим окружность талии 100 см, будет равно 30,8 кг/м².

В программе Statistica 10 предусмотрена возможность расчета 95% доверительных интервалов для среднего прогнозируемого значения зависимой переменной. Для этого потребуется вернуться к процессу анализа и с помощью нажатий на кнопку «Cancel» вернуться к окну «Multiple regression results», вкладка «Residuals/assumptions/prediction» (рисунок 12). Далее нажмем на кнопку «Predict dependent variable», после чего программа откроет окно для ввода значения независимой переменной, на основании которой будет рассчитано значение зависимой переменной и ее 95% доверительный интервал (рисунок 23). Введем значение переменной «Waist_circum», равное 100, и нажмем на кнопку «OK» для вывода результатов расчета значения зависимой переменной «BMI» (рисунок 24).

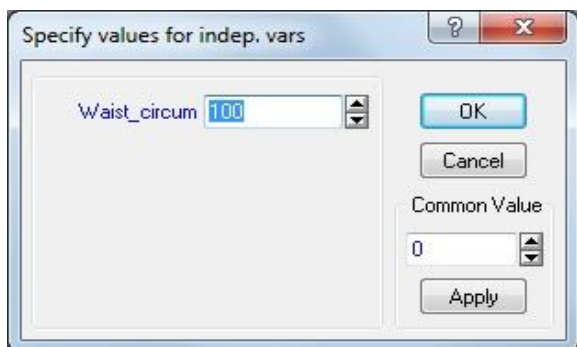


Рис. 23. Окно «Specify values for indep. vars» программы Statistica 10.

Predicting Values for (Regression_STAT variable: BMI)			
Variable	b-Weight	Value	b-Weight * Value
Waist_circum	0,393599	100,0000	39,35988
Intercept			-8,60319
Predicted			30,75669
-95,0%CL			30,39002
+95,0%CL			31,12336

Рис. 24. Результаты расчета значения ИМТ на основании значения окружности талии, равного 100 см.

Таким образом, для значения окружности талии, равного 100 см, среднее значение ИМТ пациентов с метаболическим синдромом будет с 95% вероятностью находиться в пределах от 30,39 до 31,12 кг/м² (строки «-95,0%CL» и «+95,0%CL»). Но это совершенно не значит, что при окружности талии, равной 100 см, ИМТ будет с 95% надежностью находится в указанных пределах – рассчитанный доверительный интервал относится только к среднему значению ИМТ. Соответственно, прогноз среднего значения, даже если рассчитать его 95% доверительный интервал, не имеет весомой практической ценности. С практической точки зрения гораздо важнее оценить, в каком интервале будут находиться значения ИМТ отдельных пациентов с метаболическим синдромом (разумеется, данный доверительный интервал будет шире доверительного интервала для среднего прогнозируемого значения ИМТ). Для этого необходимо рассчитать нижнюю и верхнюю границы предсказательного интервала, в который попадет 95% всех значений ИМТ пациентов из генеральной совокупности, согласно формулам:

$$\bar{Y}_i - t_{(n-2)} \cdot s_e \cdot \sqrt{1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}$$

$$\bar{Y}_i + t_{(n-2)} \cdot s_e \cdot \sqrt{1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}$$

где $\bar{Y}_i$ является спрогнозированным средним значением, рассчитанным по уравнению $Y_{BMI} = -8,603 + 0,394 \times X_{Waist_circum}$, под выражением $t_{(n-2)}$ подразумевается значение t для $n-2$ степеней свободы (поскольку выборка достаточно большая ($n = 277$), можно использовать значение 1,96, как для классического нормального распределения), а s_e представляет собой стандартное отклонение остатков. s_e можно рассчитать с использованием программных средств Statistica 10 на основании значений переменной «Residuals» таблицы, представленной на рисунке 20 («Statistics» - «Basic Statistics/Tables» - «Descriptive Statistics» - «Summary: Statistics») [15], либо вручную по формуле:

$$s_e = \sqrt{\frac{\sum_{i=1}^n (Y_i - \bar{Y}_i)^2}{n - 2}}$$

Однофакторный линейный регрессионный анализ с использованием программы SPSS 20

Для начала работы необходимо открыть файл 8_Regression_SPSS.sav, который потребуется загрузить с сайта журнала «Наука и Здоровье». В файле представлены те же вариационные ряды, что и в файле данных программы Statistica: ИМТ («BMI») и окружность талии («Waist_circum»).

На начальном этапе анализа необходимо построить скаттерограмму для подтверждения линейного характера связи между переменными.

Для этого войдем в меню «Graphs», раздел «Legacy Dialogs», подраздел «Scatter/Dot» (рисунок 25).

В открывшемся окне «Scatter/Dot» выберем простую скаттерграмму «Simple Scatter» и нажмем на кнопку «Define» (рисунок 26).

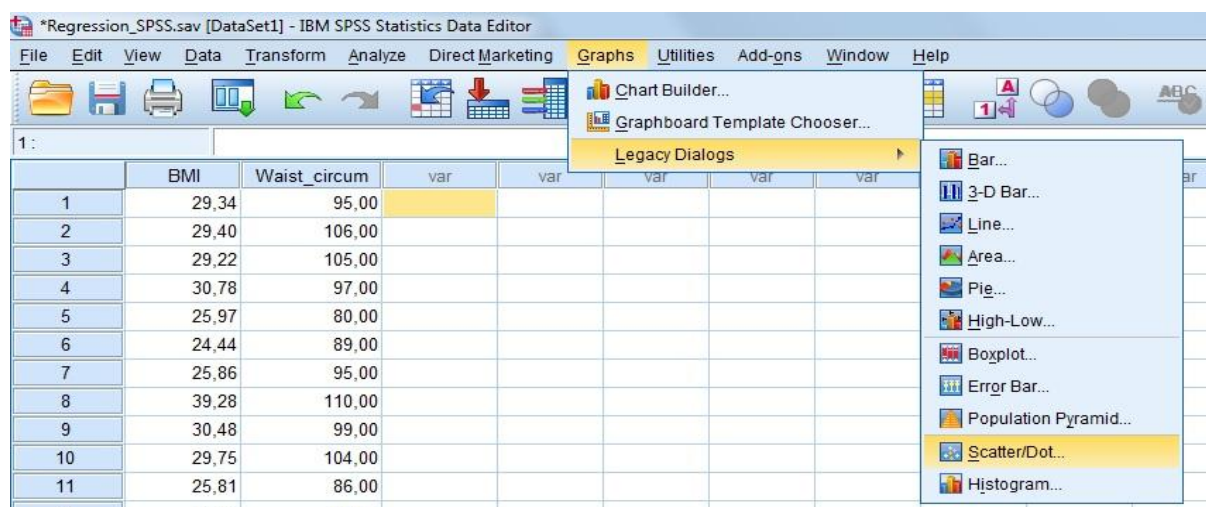


Рис. 25. Выбор подраздела «Scatter/Dot» меню «Graphs» программы SPSS 20.

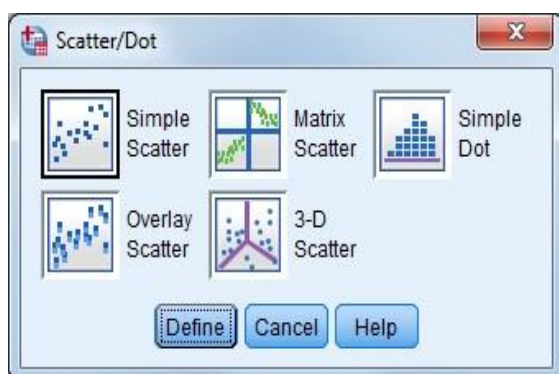


Рис. 26. Окно «Scatter/Dot» программы SPSS 20.

В открывшемся окне «Simple Scatterplot» с помощью стрелок между полями перенесем переменную «Waist_circum» в поле «X Axis», а переменную «BMI» – в поле «Y Axis» и нажмем на кнопку «OK» (рисунок 27).

В результате программа сформирует скаттерграмму, представленную на рисунке 28. Данный график соответствует скаттерграмме на рисунке 5.

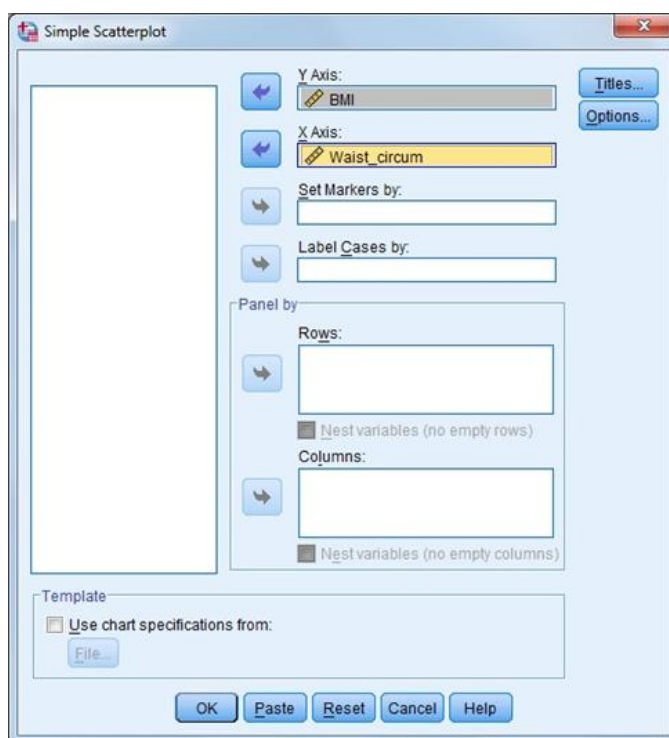


Рис. 27. Окно «Simple Scatterplot» программы SPSS 20.

Далее выполним однофакторный линейный регрессионный анализ, который в программе SPSS 20 потребует значительно меньшего количества манипуляций по сравнению с

программой Statistica 10. Для этого войдем в меню «Analyze», раздел «Regression», подраздел «Linear» (рисунок 29).

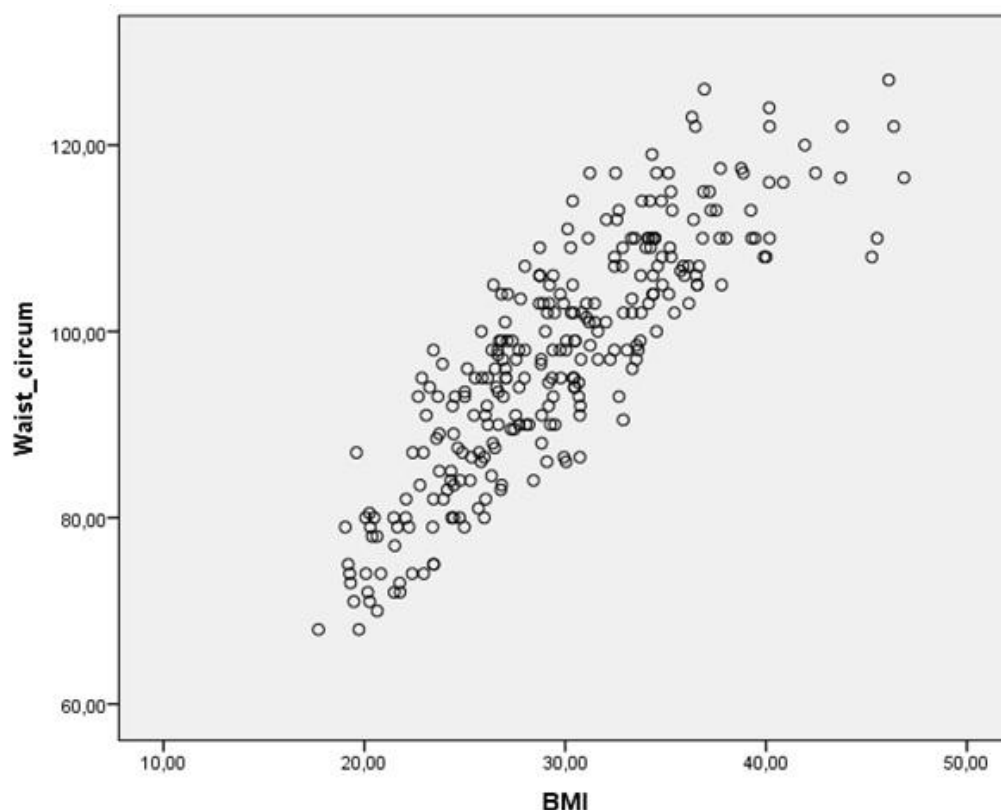


Рис. 28. Скаттерграмма зависимости переменной «BMI» от переменной «Waist_circum» (программа SPSS 20).

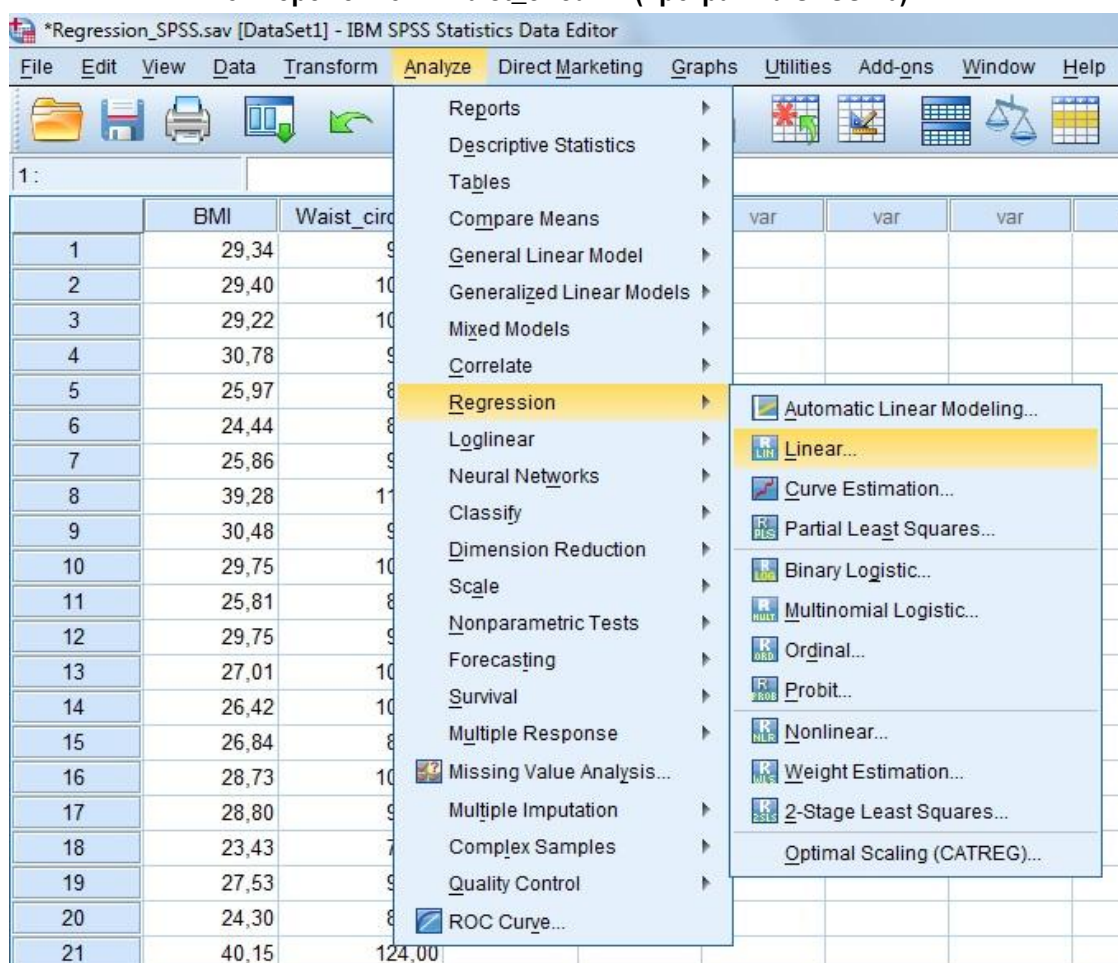


Рис. 29. Выбор подраздела «Regression» – «Linear» меню «Analyze» программы SPSS 20.

Открывается окно «Linear Regression», в котором с помощью стрелки между полями необходимо перенести переменные «BMI» и «Waist_circum» из левого поля в соответствующие правые поля для зависимой («Dependent») и независимой («Independent(s)») переменных (рисунок 30).

Далее потребуется пошагово настроить параметры вывода результатов анализа.

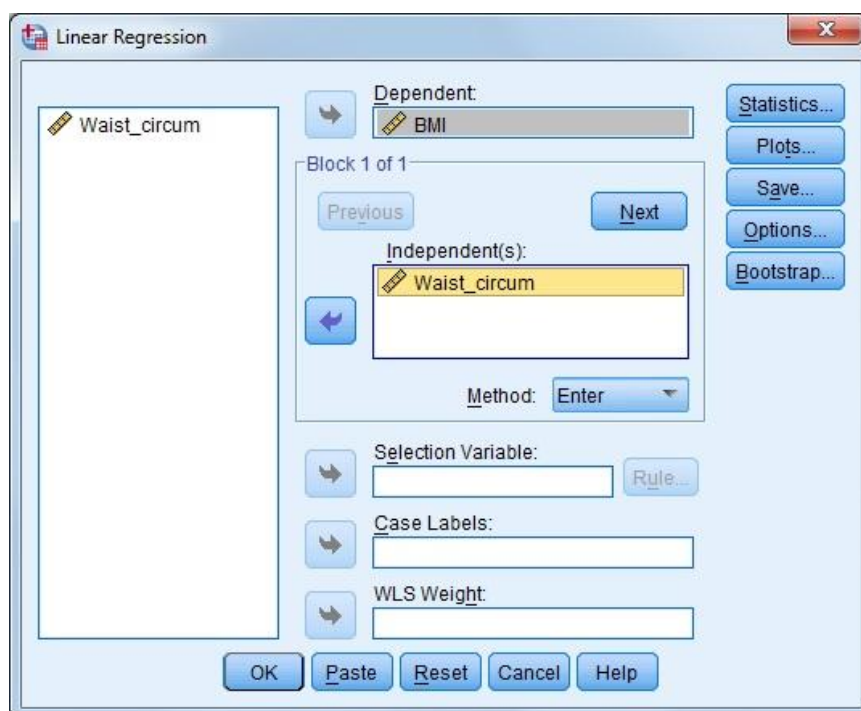


Рис. 30. Окно «Linear Regression» программы SPSS 20 (выбор зависимой переменной «BMI» и независимой переменной «Waist_circum»).

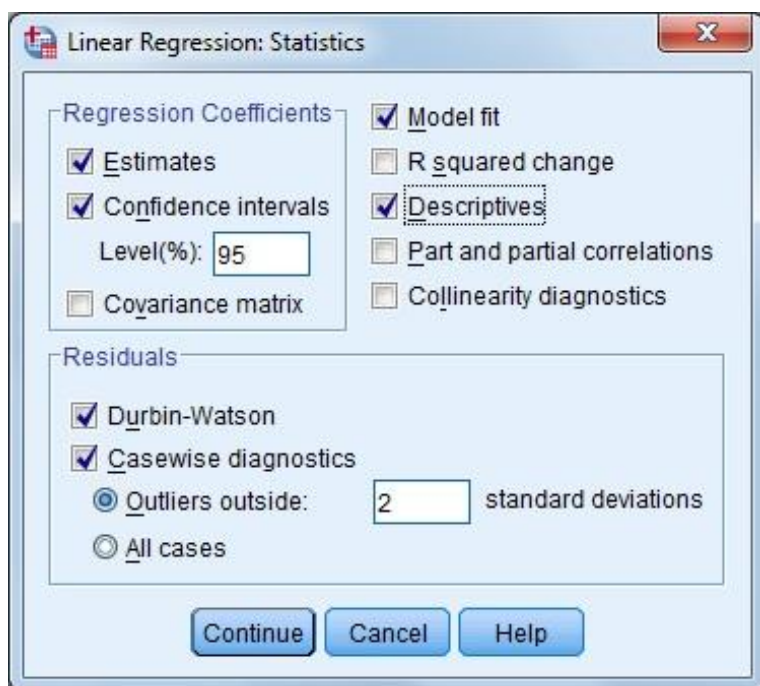


Рис. 31. Окно «Linear Regression: Statistics» программы SPSS 20.

Программа вернется к окну «Linear Regression», в котором нажмем на кнопку «Plots», и в открывшемся окне перенесем переменную «*ZPRED» в поле оси X, а переменную «*ZRESID» – в поле оси Y, после чего отметим позиции «Histogram» и «Normal probability plot» и нажмем на кнопку «Continue» (рисунок 32). Представленные действия

позволят вывести в результатах анализа скаттерграмму разброса стандартизованных остатков в зависимости от стандартизованных прогнозируемых значений зависимой переменной, а также оценить соответствие распределения остатков нормальному распределению.

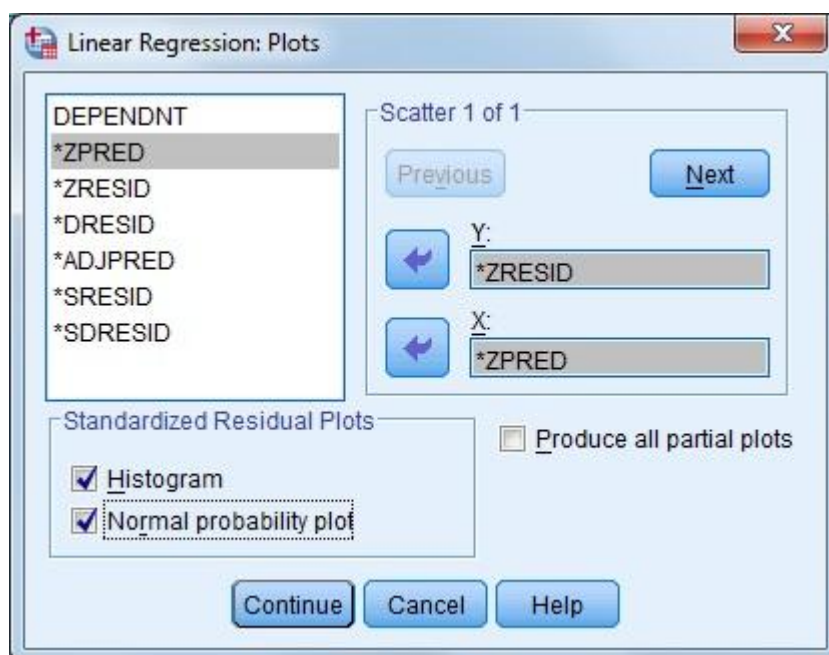


Рис. 32. Окно «Linear Regression: Plots» программы SPSS 20.

Далее в окне «Linear Regression» нажмем кнопку «Save» и отметим галочками позицию «Unstandardized» в областях «Predicted Values» и «Residuals», позиции «Mean» и «Individual» в поле «Prediction intervals», подтвердив выбор нажатием на кнопку «Continue» (рисунок 33). Эти действия создадут в исходной таблице данных новые переменные, соответствующие прогнозируемым значениям зависимой переменной и остатков для каждого наблюдения, а также границы 95% доверительного интервала для среднего прогнозируемого значения зависимой переменной и 95% доверительного интервала для значения зависимой переменной в генеральной совокупности.

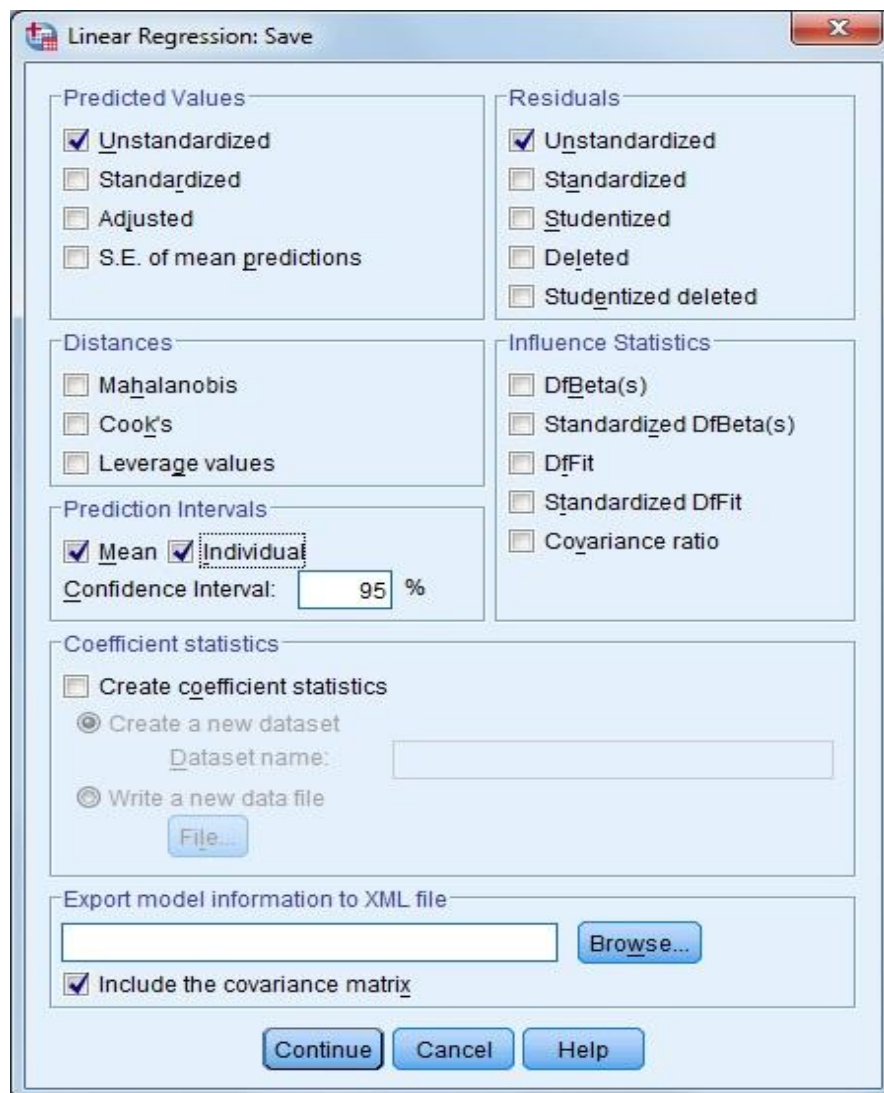


Рис. 33. Окно «Linear Regression: Save» программы SPSS 20.

Далее нажатием на кнопку «Continue» в окне «Linear Regression» запустим анализ.

Программа SPSS 20 выводит результаты моделирования в виде последовательности таблиц и графиков, основные из которых будут рассмотрены ниже. Представленные результаты однофакторного линейного регрессионного анализа не отличаются от тех, которые были описаны при рассмотрении

алгоритма работы с программой Statistica 10, и не требуют дополнительных пояснений.

В таблице 2 программа представляет значение коэффициента корреляции Пирсона для переменных «BMI» и «Waist_circum», равное 0,857 (строка «Pearson Correlation») и достигнутый уровень его статистической значимости: $p < 0,001$. (строка «Sig. (2-tailed)»).

Таблица 2.

Результаты расчета коэффициента корреляции Пирсона для переменных «BMI» и «Waist_circum».

		BMI	Waist_circum
Pearson Correlation	BMI	1,000	,857
	Waist_circum	,857	1,000
Sig. (1-tailed)	BMI	.	,000
	Waist_circum	,000	.
N	BMI	277	277
	Waist_circum	277	277

В таблице 3 представлены значения коэффициента детерминации (графа «R Square») и критерия Durbin-Watson модели (графа «Durbin-Watson»).

Таблица 3.

Значения коэффициента детерминации и критерия Durbin-Watson модели.

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,857 ^a	,735	,734	3,03856	1,935

a. Predictors: (Constant), Waist_circum

b. Dependent Variable: BMI

В таблице 4 главным является значение критерия F (графа «F») и уровень его статистической значимости (графа «Sig.»),

программа также представляет сумму квадратов модели (графа «Sum of Squares»).

Таблица 4.

Результаты проверки модели с помощью критерия F.

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7025,943	1	7025,943	760,973	,000 ^b
	Residual	2539,033	275	9,233		
	Total	9564,976	276			

a. Dependent Variable: BMI

b. Predictors: (Constant), Waist_circum

В таблице 5 программа представляет значения константы модели (графа «B», строка «(Constant)»), коэффициента регрессии (графа «B», строка «Waist_circum»), результаты проверки нулевой гипотезы об их равенстве нулю

(графы «t» и «Sig.»), а также в соответствующих строках верхнюю и нижнюю границы 95% интервала для константы и коэффициента регрессии (графа «95,0% Confidence Interval for B», подразделы «Upper Bound» и «Lower Bound» соответственно).

Таблица 5.

Значения константы и коэффициента регрессии модели.

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	-8,602	1,402		-6,136	,000	-11,362	-5,843
Waist_circum	,394	,014	,857	27,586	,000	,366	,422

a. Dependent Variable: BMI

В таблице 6 представлены результаты анализа «выбросов» (соответствуют представленным на рисунке 18).

Таблица 6.

Результаты анализа «выбросов» в выборке.

Case Number	Std. Residual	BMI	Predicted Value	Residual
14	-2,074	26,42	32,7244	-6,30085
106	3,739	45,27	33,9052	11,35999
107	3,165	46,87	37,2507	9,61716
172	3,566	45,53	34,6924	10,83694
186	-2,047	31,23	37,4475	-6,22058
243	2,125	43,71	37,2507	6,45719
247	2,008	40,01	33,9052	6,10277
263	2,285	46,36	39,4154	6,94386
268	-2,150	23,44	29,9693	-6,53180

a. Dependent Variable: BMI

Программа также выводит гистограмму распределения остатков модели (рисунок 34) и квантильную диаграмму. Данная гистограмма

распределения отличается от представленной на рисунке 16 только количеством интервалов, на которые программа разбивает выборку.

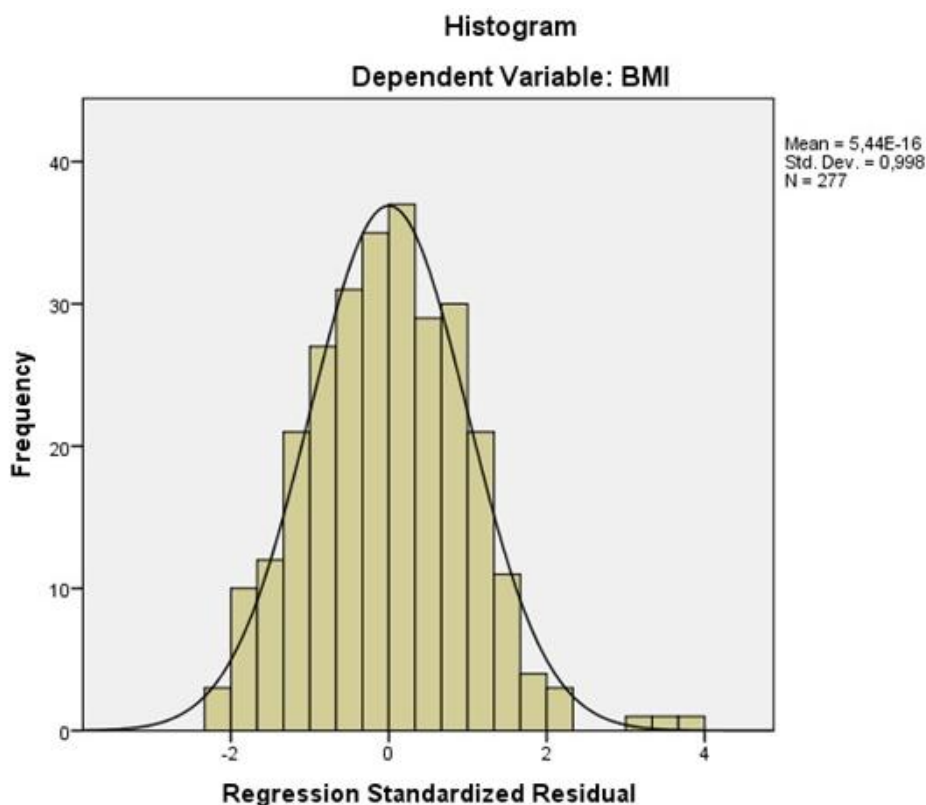


Рис. 34. Гистограмма распределения остатков модели (программа SPSS 20).

На рисунке 35 представлена значения переменной «BMI» и скаттерограмма зависимости между стандартизированными остатками, стандартизированными прогнозируемыми ответствующая представленной на рисунке 22.

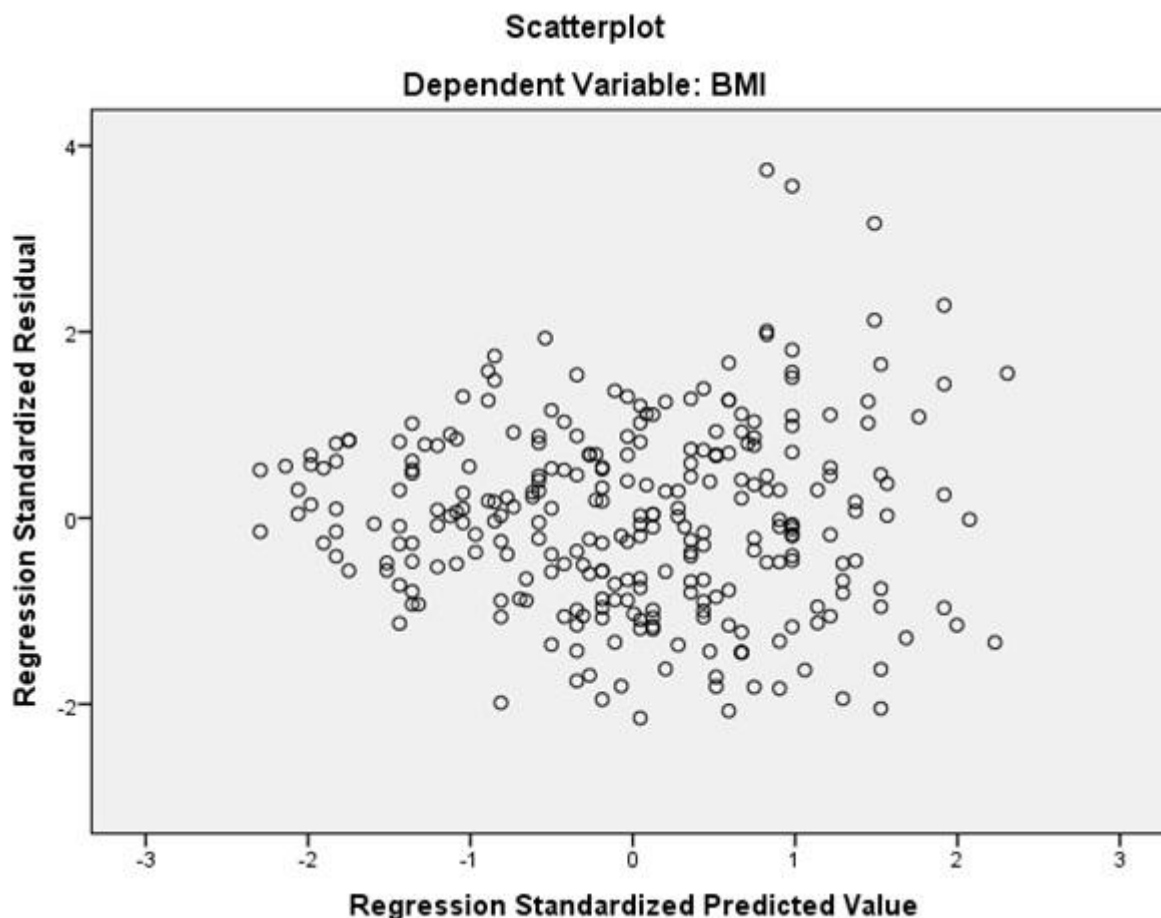
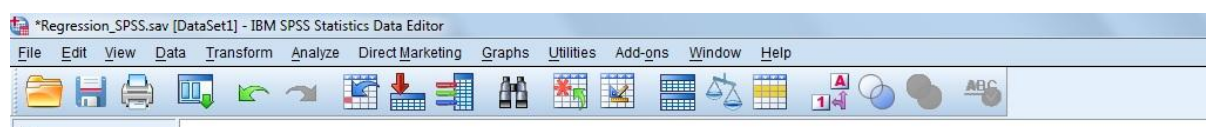


Рис. 35. Скаттерограмма зависимости между стандартизированными прогнозируемыми значениями переменной «BMI» и стандартизированными остатками (программа SPSS 20).

Следует отметить, что программа SPSS позволяет рассчитать 95% доверительные интервалы как для среднего значения зависимой переменной, так и для значения зависимой переменной в генеральной совокупности (рисунок 36). Программа создает переменные LMCI_1 и UMCI_1, показывающие нижнюю и верхнюю границы доверительного интервала для среднего прогнозируемого значения ИМТ для каждого значения окружности талии, включенного в данное

исследование. Переменные для верхней и нижней границы предсказательного интервала значения ИМТ в генеральной совокупности сохранены в виде переменных LICI_1 и UICI_1 соответственно. Например, 95% пациентов с метаболическим синдромом в генеральной совокупности с окружностью талии, равной 106 см, согласно представленным расчетам, будут иметь ИМТ от 27,12 до 39,12 кг/м² (вторая строка таблицы, представленной на рисунке 36).



	BMI	Waist_circum	PRE_1	RES_1	LMCI_1	ULCI_1	LICI_1	UICI_1
1	29,34	95,00	28,78854	,55150	28,42276	29,15431	22,79557	34,78150
2	29,40	106,00	33,11801	-3,72179	32,68526	33,55076	27,12058	39,11544
3	29,22	105,00	32,72442	-3,50388	32,30667	33,14218	26,72806	38,72079
4	30,78	97,00	29,57571	1,20519	29,21611	29,93532	23,58312	35,56831
5	25,97	80,00	22,88471	3,08541	22,27762	23,49179	16,87219	28,89723
6	24,44	89,00	26,42700	-1,98520	25,99678	26,85723	20,42976	32,42425
7	25,86	95,00	28,78854	-2,92402	28,42276	29,15431	22,79557	34,78150
8	39,28	110,00	34,69236	4,58603	34,18833	35,19640	28,68937	40,69536
9	30,48	99,00	30,36289	,12027	30,00075	30,72504	24,37014	36,35564
10	29,75	104,00	32,33083	-2,57809	31,92667	32,73500	26,33540	38,32627
11	25,81	86,00	25,24624	,56021	24,76453	25,72795	19,24508	31,24740
12	29,75	98,00	29,96930	-,21724	29,60952	30,32908	23,97670	35,96191
13	27,01	101,00	31,15007	-4,14494	30,77685	31,52329	25,15664	37,14349
14	26,42	105,00	32,72442	-6,30085	32,30667	33,14218	26,72806	38,72079
15	26,84	83,50	24,26227	2,57638	23,73121	24,79332	18,25695	30,26759
16	28,73	109,00	34,29878	-5,56703	33,81402	34,78354	28,29737	40,30018
17	28,80	97,00	29,57571	-,77130	29,21611	29,93532	23,58312	35,56831
18	23,43	75,00	20,91676	2,51716	20,19171	21,64181	14,89119	26,94234

Рис. 36. Таблица прогнозируемых на основании модели значений переменной «BMI», остатков и 95% доверительных интервалов для среднего значения переменной «BMI» и значения переменной «BMI» в генеральной совокупности.

Как видно из расчетов, разброс предсказанных индивидуальных значений достаточно велик даже при относительно высоком коэффициенте детерминации модели и относительно узком доверительном интервале для среднего. При прогнозировании индивидуальных значений зависимой переменной для тех значений независимой переменной, которые не вошли в исследование, можно воспользоваться вышеприведенными формулами для нижней и верхней границ 95% доверительного интервала для прогнозируемого значения ИМТ.

Литература:

1. Аканов А.А., Турдалиева Б.С., Изекенова А.К., Рамазанова М.А., Абдраимова Э.Т., Гржибовский А.М. Оценка использования статистических методов в научных статьях медицинских журналов Казахстана // Экология человека. 2013. №5. С. 61-64.
2. Банержи А. Медицинская статистика понятным языком: вводный курс. М. : Практическая медицина, 2007. 287 с.
3. Боровиков В. STATISTICA. Искусство анализа данных на компьютере: для профессионалов. СПб. : Питер, 2003. 688 с.
4. Бююль А., Цефель П. SPSS: искусство обработки информации. Анализ

статистических данных и восстановление скрытых закономерностей. СПб. : ООО «ДиаСофтЮП», 2005. 608 с.

5. Вараксин А.Н., Панов В.Г., Казмер Ю.И. Статистические модели с коррелированными предикторами в экологии и медицине. Екатеринбург : Изд-во Урал. ун-та. 2011. 92 с.

6. Гланц С. Медико-биологическая статистика. М. : Практика, 1998. 459 с.

7. Гржибовский А.М. Корреляционный анализ // Экология человека. 2008. №9. С. 50-60.

8. Гржибовский А.М. Однофакторный линейный регрессионный анализ // Экология человека. 2008. №10. С. 55-64.

9. Гржибовский А.М. Типы данных, проверка распределения и описательная статистика // Экология человека. 2008. №1. С. 52-58.

10. Гржибовский А.М., Иванов С.В. Анализ номинальных и ранговых переменных данных с использованием программного обеспечения Statistica и SPSS // Наука и Здравоохранение. 2016. № 6. С5-39.

11. Гржибовский А.М., Иванов С.В. Исследования типа «случай-контроль» в здравоохранении // Наука и Здравоохранение. 2015. № 4. С. 5-17

12. Гржибовский А.М., Иванов С.В. Когортные исследования в здравоохранении // Наука и Здравоохранение. 2015. № 3. С. 5-16.
13. Гржибовский А.М., Иванов С.В. Корреляционный анализ данных с использованием программного обеспечения Statistica и SPSS // Наука и Здравоохранение. 2017. № 1. С. 7-36.
14. Гржибовский А.М., Иванов С.В. Поперечные (одномоментные) исследования в здравоохранении // Наука и Здравоохранение. 2015. № 2. С. 5-18.
15. Гржибовский А.М., Иванов С.В., Горбатова М.А. Описательная статистика с использованием пакетов статистических программ Statistica и SPSS: проверка распределения // Наука и Здравоохранение. 2016. № 1. С. 7-23.
16. Гржибовский А.М., Иванов С.В., Горбатова М.А. Сравнение количественных данных двух независимых выборок с использованием программного обеспечения Statistica и SPSS: параметрические и непараметрические критерии // Наука и Здравоохранение. 2016. № 2. С. 5-28.
17. Гржибовский А.М., Иванов С.В., Горбатова М.А. Сравнение количественных данных двух парных выборок с использованием программного обеспечения Statistica и SPSS: параметрические и непараметрические критерии // Наука и Здравоохранение. 2016. № 3. С. 5-25.
18. Гржибовский А.М., Иванов С.В., Горбатова М.А. Сравнение количественных данных трех и более независимых выборок с использованием программного обеспечения Statistica и SPSS: параметрические и непараметрические критерии // Наука и Здравоохранение. 2016. № 4. С. 5-37.
19. Гржибовский А.М., Иванов С.В., Горбатова М.А. Сравнение количественных данных трех и более парных выборок с использованием программного обеспечения Statistica и SPSS: параметрические и непараметрические критерии // Наука и Здравоохранение. 2016. № 5. С. 5-29.
20. Гржибовский А.М., Иванов С.В. Экологические (корреляционные) исследования в здравоохранении // Наука и Здравоохранение. 2015. № 5. С. 5-18.
21. Гржибовский А.М., Иванов С.В. Экспериментальные исследования в здравоохранении // Наука и Здравоохранение. 2015. № 6. С. 5-17.
22. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. 3-е изд. М. : Издательский дом «Вильямс», 2007. 912 с.
23. Жунисова М.Б., Шалхарова Ж.С., Шалхарова Ж.Н., Гржибовский А.М. Типы пищевого поведения и абдоминальное ожирение // Медицина. 2015. №4. С. 92-95.
24. Жунисова М.Б., Шалхарова Ж.С., Шалхарова Ж.Н., Нускабаева Г.О., Садыкова К.Ж., Маденбай К.М., Гржибовский А.М. Психозомоциональный стресс как предиктор типа пищевого поведения в Казахстане // Экология человека. 2015. №5. С. 36-45.
25. Зайцев В.М., Лифляндский В.Г., Маринкин В.И. Прикладная медицинская статистика. СПб. : Фолиант, 2003. 428 с.
26. Зуева Л.П., Яфаев Р.Х. Эпидемиология : учебник. СПб : ООО «Издательство Фолиант», 2008. 752 с.
27. Лакин Г.Ф. Биометрия. М. : Высшая школа, 1990. 351 с.
28. Маденбай К.М., Шалхарова Ж.С., Шалхарова Ж.Н., Жунисова М.Б., Садыкова К.Ж., Нускабаева Г.О., Гржибовский А.М. Оценка связи между площадью подкожной жировой ткани и показателями электронейромиографии // Экология человека. 2015. №7. С. 58-64.
29. Петри А., Сэбин К. Наглядная статистика в медицине. М. : ГЭОТАР-Мед, 2003. 140 с.
30. Рахымбеков Т.К., Гржибовский А.М. К вопросу о необходимости повышения качества казахстанских научных публикаций для успешной интеграции в международное научное сообщество // Наука и Здравоохранение. 2015. №1. С. 5-11.
31. Реброва О.Ю. Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA. М. : МедиаСфера, 2002. 312 с.
32. Садыкова К.Ж., Шалхарова Ж.С., Нускабаева Г.О., Садыкова А.Д., Жунисова М.Б., Маденбай К.М., Гржибовский А.М. Распространенность анемии, ее социально-демографические детерминанты и возможная связь с метаболическим синдромом в г.

Туркестан, Южный Казахстан // Экология человека. 2015. №8. С. 58-64.

33. Субботина А.В., Гржибовский А.М. Описательная статистика и проверка нормальности распределения количественных данных // Экология человека. 2014. №2. С. 51-57.

34. Унгурияну Т.Н., Гржибовский А.М. Корреляционный анализ с использованием пакета статистических программ STATA // Экология человека. 2014. №9. С. 60-64.

35. Флетчер Р. Клиническая эпидемиология. Основы доказательной медицины: пер. с англ. / Р. Флетчер, С. Флетчер, Э. Вагнер. М. : Медиа Сфера, 1998. 352 с.

36. Юнкеров В.И., Григорьев С.Г. Математико-статистическая обработка данных медицинских исследований. СПб :ВМедА, 2002. 266 с.

37. Beaglehole R., Bonita R. Basic epidemiology. World Health Organization, Geneva, 1993.

38. Cleopas T.J. et al. Statistics Applied to Clinical Trials. 4th ed. Springer, 2009.

References:

1. Akanov A.A., Turdalieva B.S., Izenkova A.K., Ramazanov M.A., Abdaimova, Grjibovski A.M. Otsenka ispol'zovania statisticheskikh metodov v nauchnih statyakh Kazakhstana [Assesment of use of statistical methods in scientific articles of the Kazakhstan's medical journals]. *Ekologiya cheloveka* [Human Ecology]. 2013, No.5, pp. 61-64. [in Russian]

2. Banerzhi A. Meditsinskaya statistika ponyatnym yazykom : vvodnyy kurs [Medical statistics in plain language : an introductory course]. M. : *Prakticheskaya meditsina* , 2007.. 287 p. [in Russian]

3. Borovikov V. STATISTICA. Iskusstvo analiza dannykh na kompyutere: dlya professionalov [STATISTICA. The art of data analysis using computer: for professionals]. SPb. : Piter, 2003. p. 688.

4. Buhl A., Zofel P. SPSS: isskustvo obrabotki informatsii. Analiz statisticheskikh daanikh i vosstanovlenie skritikh zakonomernostey [SPSS: the art of information analysis. Statistical data analysis and hidden regularities identification]. SPb. : ООО «DiaSoftUP», 2005. 608 p. [in Russian]

5. Varaksin A.N., Panov V.G., Kazmer Yu.I. *Statisticheskiye modeli s korrelirovannymi prediktorami v ekologii i meditsine* [Statistical models with correlated predictors in ecology and medicine]. Yekaterinburg : Izd - vo Ural. un-ta. 2011. 92 p. [in Russian]

6. Glants S. *Mediko-biologicheskaya statistika* [The biomedical statistics]. M. : Praktika, 1998. p. 459. [in Russian]

7. Grjibovski A.M. Korrelatsionniy analiz [Correlation analysis]. *Ekologiya cheloveka* [Human Ecology]. 2008, No.9, pp. 50-60. [in Russian]

8. Grjibovski A.M. Odnofaktornyy lineynyy regressionnyy analiz [Simple linear regression analysis]. *Ekologiya cheloveka* [Human Ecology]. 2008, No.10, pp. 55-64. [in Russian]

9. Grjibovski A.M. Tipy dannykh, proverka raspredeleniya i opisatel'naya statistika [Types of data, distribution estimation and descriptive statistics]. *Ekologiya cheloveka* [Human Ecology]. 2008, No.1, pp. 52-58. [in Russian].

10. Grjibovski A.M., Ivanov S.V. Analiz nominalnykh i rangovykh peremennykh dannykh s ispol'zovaniyem programmnoy obespecheniya Statistica i SPSS [Analysis of nominal and ordinal data using Statistica and SPSS software]. *Nauka i Zdravoohranenie* [Science & Healthcare]. 2016, № 6, pp. 5-39. [in Russian].

11. Grjibovski A.M., Ivanov S.V. Issledovaniya tipa sluchay-kontrol v zdravoohranenii [Case-control studies in health sciences]. *Nauka i Zdravoohranenie* [Science & Healthcare]. 2015, 4, pp. 5-17 [in Russian].

12. Grjibovski A.M., Ivanov S.V. Kogortnie issledovaniya v zdravoohranenii [Cohort studies in health sciences]. *Nauka i Zdravoohranenie* [Science & Healthcare]. 2015, 3, pp. 5-16. [in Russian].

13. Grjibovski A.M., Ivanov S.V. Korrelyatsionnyy analiz dannykh s ispol'zovaniyem programmnoy obespecheniya Statistica i SPSS [Correlation analysis using software Statistica and SPSS]. *Nauka i Zdravoohranenie* [Science & Healthcare]. 2017, 1, PP. 7-36. [in Russian]. [in Russian].

14. Grjibovski A.M., Ivanov S.V. Poperechnie (odnomomentnie) issledovaniya v zdravoohranenii [Cross-sectional studies in health sciences]. *Nauka i Zdravoohranenie* [Science & Healthcare]. 2015, No2, PP. 5-18. [in Russian]

15. Grijbovski A.M., Ivanov S.V., Gorbatova M.A. Opisatel'naya statistika s ispol'zovaniyem paketov statisticheskikh programm Statistica i SPSS I proverka raspredeleniya [Descriptive statistics using Statistica and SPSS software]. *Nauka i Zdravookhranenie* [Science & Healthcare]. 2016, 1, pp. 7-23 [in Russian].
16. Grijbovski A.M., Ivanov S.V., Gorbatova M.A. Sravneniye kolichestvennykh dannyykh dvukh nezavisimyykh vyborok s ispol'zovaniyem programmogo obespecheniya Statistica i SPSS : parametricheskiye i neparametricheskiye kriterii [Comparing the quantitative data of two independent groups using the software Statistica and SPSS: parametric and nonparametric tests]. *Nauka i Zdravookhranenie* [Science & Healthcare]. 2016, 2, pp.5-28 [in Russian].
17. Grijbovski A.M., Ivanov S.V., Gorbatova M.A. Sravneniye kolichestvennykh dannyykh dvukh parnykh vyborok s ispol'zovaniyem programmogo obespecheniya Statistica i SPSS : parametricheskiye i neparametricheskiye kriterii [Comparing the quantitative data of two dependent variations using the software Statistica and SPSS: parametric and nonparametric tests]. *Nauka i Zdravookhranenie* [Science & Healthcare]. 2016, 3, pp. 5-25. [in Russian].
18. Grijbovski A.M., Ivanov S.V., Gorbatova M.A. Sravneniye kolichestvennykh dannyykh trekh i boleye nezavisimyykh vyborok s ispol'zovaniyem programmogo obespecheniya Statistica i SPSS : parametricheskiye i neparametricheskiye kriterii [Comparing of the quantitative data of three or more independent samples using Statistica and SPSS software: parametric and nonparametric methods]. *Nauka i Zdravookhranenie* [Science & Healthcare]. 2016, 4, pp. 5-37. [in Russian].
19. Grijbovski A.M., Ivanov S.V., Gorbatova M.A. Sravneniye kolichestvennykh dannyykh trekh i boleye parnykh vyborok s ispol'zovaniyem programmogo obespecheniya Statistica i SPSS : parametricheskiye i neparametricheskiye kriterii [Comparing of the quantitative data of three or more dependent samples using Statistica and SPSS software: parametric and nonparametric methods]. *Nauka i Zdravookhranenie* [Science & Healthcare]. 2016, 5, pp. 5-29. [in Russian].
20. Grijbovski A.M., Ivanov S.V. Ekologicheskkiye (korrelyatsionnyye) issledovaniya v zdravookhraneni. [Ecological (correlational) studies in health sciences]. *Nauka i Zdravookhranenie* [Science & Healthcare]. 2015, 5, pp. 5-18. [in Russian].
21. Eksperimentalnye issledovaniya v zdravookhraneni [Experimental studies in health sciences]. *Nauka i Zdravookhranenie* [Science & Healthcare]. 2015, 6, pp. 5-17. [in Russian].
22. Dreyper N., Smit G. *Prikladnoy regressionnyy analiz* [Applied regression analysis]. 3-ye izd . M. : Izdatel'skiy dom "Vil'yams", 2007. 912 p.
23. Zhunissova M.B., Shalkarova Zh.S., Shalkarova Zh. N., Nuskabayeva G.O., Sadykova K. Zh., Madenbay K.M., Grijbovski A.M. Tipy pischevogo povedeniya i abdominalnoe ozhirenie [Eating behavior types and abdominal obesity]. *Meditsina* [Medicine]. 2015, No.4. pp. 92-95. [in Russian].
24. Zhunissova M.B., Shalkarova Zh.S., Shalkarova Zh. N., Nuskabayeva G.O., Sadykova K. Zh., Madenbay K.M., Grijbovski A.M. Psykhoemotsionalniy stress kak predictor tipa pischevogo povedeniya v Kazakhstane [Psychoemotional stress and eating behavior in Kazakhstan]. *Ekologiya cheloveka* [Human Ecology]. 2015, No.5. pp. 36-45. [in Russian]
25. Zaytsev V.M., Lifyandskiy V.G., Marinkin V.I. *Prikladnaya meditsinskaya statistika* [Applied medical statistics] . SPb . : Foliant , 2003. P. 428. [in russian]
26. Zueva L.P., Yafaev R.H. *Epidemiologiya: uchebnik* [Epidemiology: the textbook]. SPb : OOO «Izdatelstvo Foliant», 2008. P. 752. [in Russian].
27. Lakin G.F. *Biometria* [Biometrics]. M. : Viscshaya shkola. 1990. P. 351. [in Russian].
28. Madenbay K.M., Shalkarova Zh.S., Shalkarova Zh. N., Zhunissova M.B., Sadykova K. Zh., Nuskabayeva G.O., Grijbovski A.M. Otsenka svyazi mezhdu ploshadyu podkojnoj jirovoy tkani i pokazatelyami electroneymyografi [Assesment of the relationship between subcutaneous fat tissue an results of electromyoneurography]. *Ekologiya cheloveka* [Human Ecology]. 2015, No.7, pp. 58-64. [in Russian].
29. Petri A., Sebin K. *Naglyadnaya statistika v meditsine* [Demonstrative statistics in medicine]. M. : GEAOTAR-Med, 2003. p. 140. [in Russian].
30. Rakhypbekov T.K., Grijbovski A.M. K voprosu o nedhodimosti povisheniya kachestva

kazakhstanskikh nauchih publicatsiy dlya uspezhnoy integratsii v mezhdunarodnoe nauchnoe soobshchestvo [The need for improvement of the quality of Kazakhstani publications for successful integration in the international research community]. *Nauka i Zdravoohranenie* [Science & Healthcare]. 2015, No.1, pp. 5-11 [in Russian].

31. Rebrova O. Yu. *Statisticheskiy analiz meditsinskih daanikh. Primenenie paketa prikladnikh program STATISTICA* [Statistical analysis of medical data. Using of STATISTICA software]. M. : MediaSphera, 2002. P. 312 [in Russian].

32. Sadykova K.Zh., Shalkharova Zh.S., Shalkharova Zh.N. Nuskabaeva G.O., Sadykova A.D., Zhunisova M.B., Madenbay K.M., Grijbovski A.M. Rasprostranennost' anemii , yeye sotsial'no - demograficheskiye determinanty i vozmozhnaya svyaz' s metabolicheskim sindromom v g. Turkestan , Yuzhniy Kazakhstan [Prevalence of anemia, its socio-demographic determinants and potential association with metabolic syndrome in residents of Turkestan, Southern Kazakhstan]. *Ekologiya cheloveka* [Human Ecology]. 2015, No.8, pp. 58-64. [in Russian]

33. Subbotina A.V., Grijbovski A.M. Opisatel'naya statistika i proverka normal'nosti raspredeleniya kolichestvennih dannikh [Descriptive statistics and normality testing for quantitative data]. *Ekologiya cheloveka* [Human Ecology]. 2014, No.2, pp. 51-57. [in Russian].

34. Unguryanu T.N., Grijbovski A.M. Korrelatsionniy analiz s ispol'zovaniyem paketa statisticheskikh programm STATA [Correlation analysis using STATA] *Ekologiya cheloveka* [Human Ecology]. 2014, No.9, pp. 60-64. [in Russian].

35. Fletcher R. et al. *Klinicheskaya epidemiologiya. Osnovy dokazatel'noi meditsiny* [Clinical epidemiology. Basics of the evidence-based medicine] / R. Fletcher, C. Fletcher, E. Vagner. M.: Media Sphere, 1998. 352 p. [in Russian].

36. Yunkerov V.I., Grigoryev S.G. *Matematiko-statisticheskaya obrabotka daanikh meditsinskih issledovaniy* [Mathematical and statistical analysis of medical research data]. SPb : VMedA, 2002. P. 266. [in Russian]

37. Beaglehole R., Bonita R. *Basic epidemiology*. World Health Organization, Geneva, 1993.

38. Cleopas T.J. et al. *Statistics Applied to Clinical Trials*. 4th ed. *Springer*, 2009.

Контактная информация:

Грибовский Андрей Мечиславович – доктор медицины, магистр международного общественного здравоохранения, Старший советник Национального Института Общественного Здравоохранения, г. Осло, Норвегия; Заведующий ЦНИЛ СГМУ, г. Архангельск, Россия; Профессор Северо-Восточного Федерального Университета, г. Якутск, Россия; Профессор, Почетный доктор Международного Казахско-Турецкого Университета им. Х.А. Ясави, г. Туркестан, Казахстан; Почетный профессор ГМУ г. Семей, Казахстан.

Почтовый адрес: INFA, Nasjonalt folkehelseinstitutt, Postboks 4404 Nydalen, 0403 Oslo, Norway.

Email: Andrej.Grijbovski@gmail.com

Телефон: +4745268913 (Норвегия), +79214717053 (Россия), +77471262965 (Казахстан).