Получена: 5 мая 2017 / Принята: 18 мая 2017 / Опубликована online: 30 июня 2017

УДК 614.2 + 303.4

ПРИМЕНЕНИЕ МНОЖЕСТВЕННОГО ЛИНЕЙНОГО РЕГРЕССИОННОГО АНАЛИЗА В ЗДРАВООХРАНЕНИИ С ИСПОЛЬЗОВАНИЕМ ПАКЕТА СТАТИСТИЧЕСКИХ ПРОГРАММ SPSS

Екатерина Е. Шарашова [1]
Камила К. Холматова [2]
Мария А. Горбатова [2], **http://orcid.org/0000-0002-6363-9595**
Андрей М. Гржибовский [2-5], **http://orcid.org/0000-0002-5464-0498**

[1] **Арктический университет Норвегии, Тромсё, Норвегия;**
[2] **Северный Государственный Медицинский Университет, г. Архангельск, Россия;**
[3] **Национальный Институт Общественного Здравоохранения, г. Осло, Норвегия;**
[4] **Международный Казахско-Турецкий Университет им. Х.А. Ясави, г. Туркестан, Казахстан;**
[5] **Северо-Восточный Федеральный Университет, г. Якутск, Россия.**

Резюме

В данной статье представлены теоретические основы проведения множественного линейного регрессионного анализа для прогнозирования значения одной зависимой количественной переменной на основании нескольких независимых при использования пакета прикладных статистических программ SPSS, описаны принципы интерпретации полученной информации на практическом примере, а также обозначены основные проблемы, возникающие при использовании этого метода и предложены варианты их решения.

Ключевые слова: множественный линейный регрессионный анализ, коэффициент детерминации, метод наименьших квадратов, доверительные интервалы, SPSS.

Abstract

APPLICATION OF THE MULTIVARIABLE LINEAR REGRESSION ANALYSIS IN HEALTHCARE USING SPSS SOFTWARE

Ekaterina E. Sharashova [1]
Kamila K. Kholmatova [2]
Maria A. Gorbatova [2], **http://orcid.org/0000-0002-6363-9595**
Andrej M. Grjibovski [2-5], **http://orcid.org/0000-0002-5464-0498**

[1] **Arctic University of Norway, Tromsø, Norway;**
[2] **Northern State Medical University, Arkhangelsk, Russia;**
[3] **Norwegian Institute of Public Health, Oslo, Norway;**
[4] **International Kazakh-Turkish University, Turkestan, Kazakhstan;**
[5] **North-Eastern Federal University, Yakutsk, Russia.**

In this article we present theoretical basis for conducting multivariable linear regression analysis for predicting the one dichotomous outcome based on several independent variables using the SPSS software. The article describes the principles of interpretation of the results using practical examples. We also describe advantages and disadvantages of this type of analysis

Key words: multivariable linear regression analysis, coefficient of determination, Least squares distance method, confidence intervals, SPSS.

Түйіндеме

SPSS СТАТИСТИКАЛЫҚ БАҒДАРЛАМАЛАР ПАКЕТІН ПАЙДАЛАНУМЕН ДЕНСАУЛЫҚ САҚТАУДАҒЫ КӨПШІЛІК СЫЗЫҚТЫҚ РЕГРЕССИВТІК ТАЛДАУДЫ ҚОЛДАНУ

Екатерина Е. Шарашова [1]
Камила К. Холматова [2]
Мария А. Горбатова [2], **http://orcid.org/0000-0002-6363-9595**
Андрей М. Гржибовский [2-5], **http://orcid.org/0000-0002-5464-0498**

[1] **Норвегия Арктикалық университеті, Тромсё, Норвегия;**
[2] **Солтүстік Мемлекеттік Медициналық Университеті, Архангельск қ., Ресей;**
[3] **Қоғамдық Денсаулық сақтау Ұлттық Институты, Осло қ., Норвегия;**
[4] **Х.А. Ясави ат. Халықаралық Қазақ – Түрік Университеті, Туркестан, Қазақстан;**
[5] **Солтүстік - Шығыс Федералдық Университеті, Якутск қ., Ресей;**

Осы мақалада SPSS қолданбалы статистикалық бағдарламаларды пайдалану кезіндегі бірнеше тәуелсіздер негізінде бір тәуелді сандық ауыспалының мәнін болжау үшін көпшілік сызықтық регрессивтік талдауды өткізудің теориялық негіздері берілген, тәжірибелік мысалда алынған ақпаратты интерпретациялары принциптері көрсетілген, сол сияқты осы әдісті қолдану кезінде шыққан негізгі мәселелер анықталды және оларды шешудің нұсқалары ұсынылған.

Негізгі сөздер: көпшілік сызықтық регрессивтік талдау, детерминация коэффициенті, ең аз квадраттар әдісі, сенімділік интервалдары, SPSS.

Библиографическая ссылка:

Шарашова Е.Е., Холматова К.К., Горбатова М.А., Гржибовский А.М. Применение множественного линейного регрессионного анализа в здравоохранении с использованием пакета статистических программ SPSS / / Наука и Здравоохранение. 2017. №3. С. 5-31.

Sharashova E.E., Kholmatova K.K., Gorbatova M.A., Grjibovski A.M. Application of the multivariable linear regression analysis in healthcare using SPSS software. *Nauka i Zdravookhranenie* [Science & Healthcare]. 2017, 3, pp. 5-31.

Шарашова Е.Е., Холматова К.К., Горбатова М.А., Гржибовский А.М. SPSS статистикалық бағдарламалар пакетін пайдаланумен денсаулық сақтаудағы көпшілік сызықтық регрессивтік талдауды қолдану / / Ғылым және Денсаулық сақтау. 2017. №3. Б. 5-31.

В наших предыдущих публикациях [1, 2] были описаны теоретические принципы проведения однофакторного линейного регрессионного анализа, применяемого в случаях, когда с помощью значения одной независимой количественной переменной (предиктора) требуется прогнозировать значение одной зависимой количественной переменной (переменной отклика). При этом данные переменные должны иметь между собой линейную зависимость. В результате проведения однофакторного регрессионного анализа можно оценить степень и определить направление линейной связи между этими количественными переменными.

В практической деятельности чаще всего требуется изучить влияние не одного, а сразу нескольких (двух и более) предикторов на переменную отклика. В данной ситуации следует использовать множественный линейный регрессионный анализ (multiple linear regression). Эта более сложная разновидность линейного регрессионного анализа позволяет не только предсказывать значение независимой переменной по известным значениям нескольких переменных-предикторов, но также оценить степень независимого друг от друга влияния каждого из предикторов на значение переменной отклика.

Суммируем теоретические основы линейного регрессионного анализа [3-5].

Основной смысл линейной регрессии состоит в том, чтобы предсказать значение зависимой переменной (Y_i) по известным значениям одной или нескольких независимых переменных (X_i), используя общее уравнение:

$$Y_i = (\text{модель}_i) + \text{ошибка}_i.$$

Значение переменной отклика, которое мы пытаемся предсказать для определенного индивидуума (Y_i), может быть выявлено с помощью определенной модели с учетом некоторой ее неточности, или случайной ошибки (ε_i). В линейном регрессионном анализе модель является линейной и для простой линейной регрессии представляет собой уравнение прямой линии:

$$Y_i = (b_0 + b_1 \cdot X_i) + \varepsilon_i,$$

где Y_i – значение зависимой переменной,

X_i – значение независимой переменной,

b_0 – константа,

b_1 – регрессионный коэффициент,

ε_i – случайная ошибка.

Для того чтобы построить это уравнение (найти коэффициенты b_0 и b_1), необходимо измерить значения зависимой (Y) и независимой (X) переменных у ряда индивидуумов (обследовать определенную выборку, получить фактические данные значений X и Y). На основании фактических данных с помощью метода наименьших квадратов создается простая регрессионная модель – уравнение прямой линии, которая наилучшим образом описывает собранные данные. Суть метода наименьших квадратов состоит в том, что из множества возможных линий (моделей) выбирается та, которая наиболее точно соответствует собранным данным. В результате получается такая линейная модель, для которой сумма квадратов различий между этой моделью - прямой линией (предсказываемые значения, Y^) и имеющимися актуальными значениями зависимой переменной, полученными на выборке (наблюдаемые значения, Y) минимизируется. Эта по возможности минимизированная неточность модели и отражена в уравнении с помощью случайной ошибки (ε_i). Те же самые теоретические принципы лежат и в основе множественного линейного регрессионного анализа с той лишь разницей, что последний используется в ситуациях с несколькими независимыми переменными. Вследствие этого, модель, отражающая линейную взаимосвязь между единственной зависимой и несколькими независимыми переменными (также прямая линия), несколько усложняется:

$$Y_i = (b_0 + b_1 \cdot x_{1i} + b_2 \cdot x_{2i} + \cdots + b_n \cdot X_{ni}) + \varepsilon_i,$$

где Y_i – значение зависимой переменной,

X_1, X_2, $\cdots$ X_n – значения первой, второй, n-ой независимых переменных,

b_0 – константа,

b_1, b_2, $\cdots$ b_n– регрессионные коэффициенты для соответствующих переменных,

ε_i – разница между предсказываемым и фактическим значением зависимой переменной Y для i-ого участника (случайная ошибка модели).

По существу, это то же уравнение прямой линии, что и в простом регрессионном анализе, с той лишь разницей, что для второго и каждого следующего предиктора, включаемого в модель, добавляется собственный регрессионный коэффициент, а переменная отклика зависит от комбинации произведений значений каждого из предикторов и соответствующих коэффициентов регрессии плюс случайная ошибка модели. Визуально представить эту линию несколько сложнее, чем в простой регрессии, т.к. она ориентирована в трех-, четырех- и т.д. мерном пространстве (в зависимости от количества переменных, включенных в модель), а не в плоскости (двухмерном пространстве), как в простой регрессионной модели с зависимой и независимой переменными.

Все это выглядит весьма абстрактно, так что давайте рассмотрим проведение множественного линейного регрессионного анализа на примере Северодвинского когортного исследования. В ходе этого исследования в 1999 году в Северодвинске (Северо-Запад России) на выборке из 869 первородящих женщин, имевших одноплодную беременность и срочные роды, были получены среди прочих данные по возрасту (полных лет) – переменная «vozrast», гестационному сроку – переменная «srok», полу ребенка «pol», а также длине «dlina» и массе тела «ves» ребенка при рождении. Более подробно дизайн и результаты данного

исследования были описаны ранее [6-9]. Результаты однофакторного регрессионного анализа с целью предсказания массы новорожденного по известному значению его длины при рождении на примере данного исследования уже были представлены [1], однако, мы приведем часть этой информации в данной статье при описании этапов множественной линейной регрессии. Предположим, что такие переменные, как длина при рождении, гестационный срок, возраст матери и др. могут оказывать какое-то влияние на массу тела новорожденного. Множественный линейный регрессионный анализ позволяет нам включить в модель в качестве переменных-предикторов все интересующие нас показатели, которые теоретически также могут влиять на вес при рождении. Мы определим, какое влияние они оказывают на массу тела ребенка независимо друг от друга, т.е. объясняет ли каждый из них сам по себе какую-то долю вариабельности зависимой переменной (массы ребенка).

Прежде чем выполнять множественный регрессионный анализ с помощью пакета прикладных статистических программ SPSS, мы можем посмотреть характер взаимосвязи между интересующими нас переменными. Используя простую двухмерную скаттерограмму, мы можем посмотреть взаимосвязь зависимой переменной с каждой из независимых по отдельности, определив при этом, носит она линейный характер или нет.

Для построения скаттерограммы в выпадающем меню «Graphs» следует выбрать окно «Scatter/Dot» (рис. 1), далее «Simple Scatter», путем нажатия на «Define».

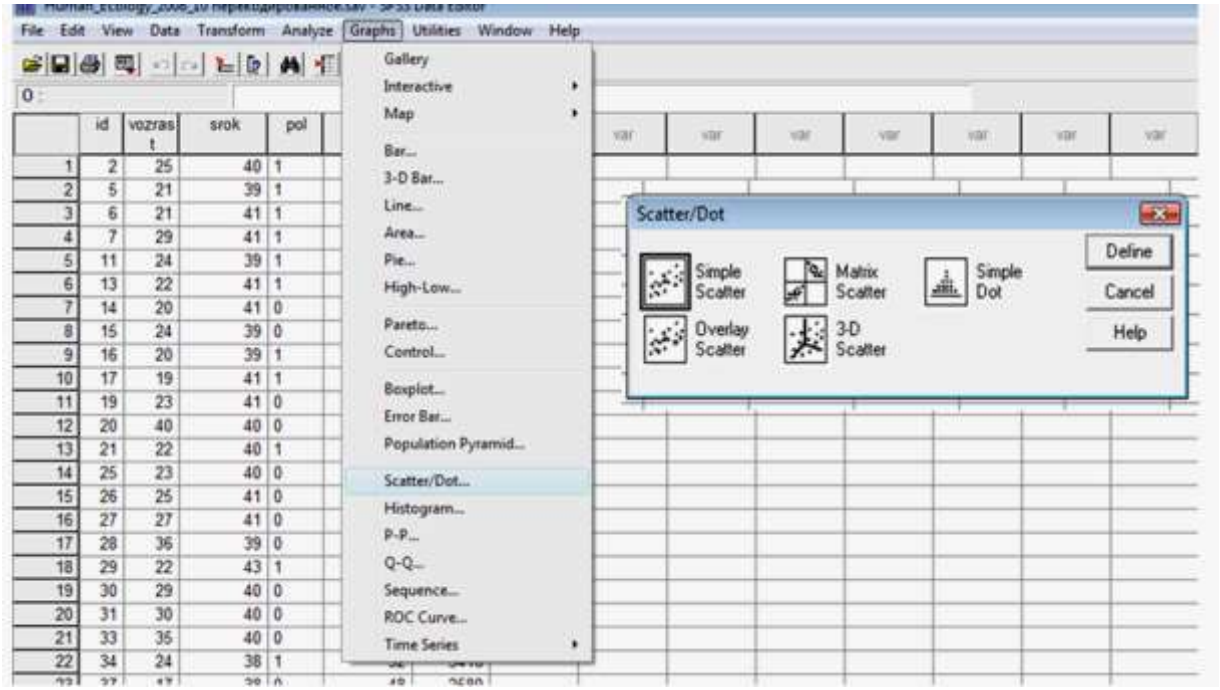

Рисунок 1. Окно «Scatter/Dot».

В появившемся окне «Simple Scatterplot» (рис. 2) можно перемещать переменные из общего левого поля в одно из правых полей, в зависимости от того какую переменную по какой из осей Вы хотели бы разместить. В представленном примере мы внесли длину тела новорожденных в поле оси абсцисс, а переменную «ves» в окно оси ординат. Подобным образом следует поступить с каждым из оставшихся предикторов. Получившиеся скаттерограммы изображены на рис. 3.

На рисунке 3 вверху изображена скаттерограмма взаимосвязи массы тела новорожденных и длины тела при рождении, из которой видно, что существует линейная связь между переменной «ves» и переменной «dlina» (R^2=0,694).

Внизу слева изображена скаттерограмма взаимосвязи массы тела с возрастом матери. На ней видно, что линейная взаимосвязь между этими переменными существует, хотя и слабо выражена: R^2=0,003.

Похожая картина наблюдается и в случае зависимости между массой тела и гестационным сроком новорожденных (рис. 3, внизу справа).

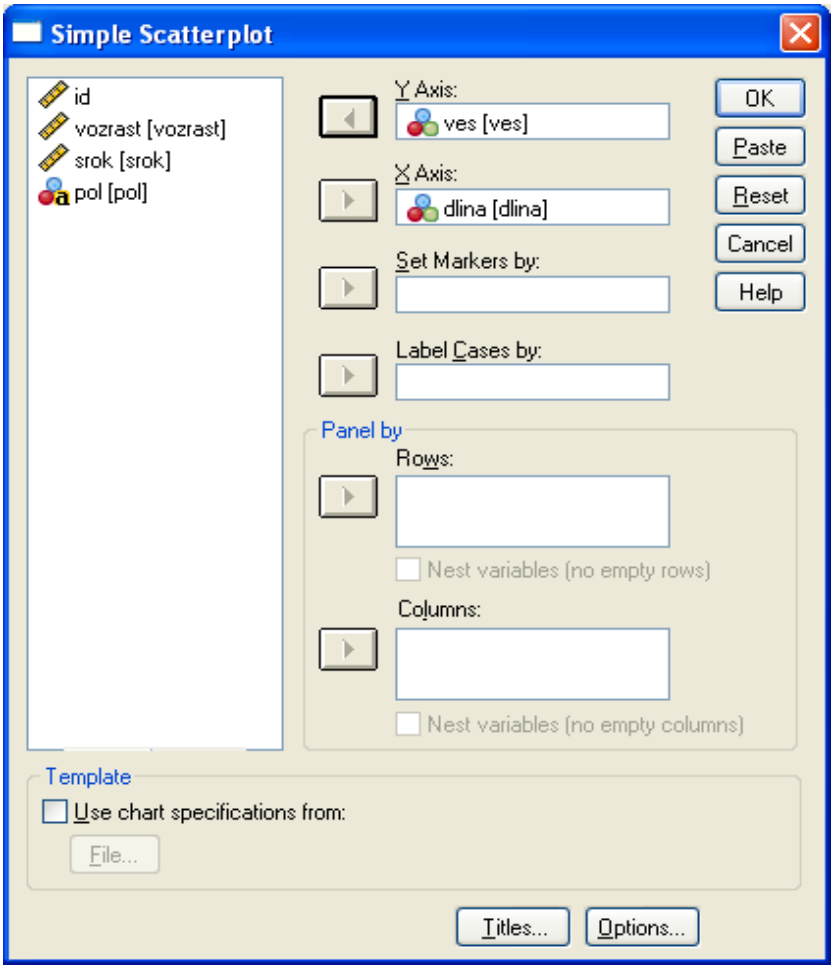

Рисунок 2. Окно «Simple Scatterplot».

Она также носит линейный характер, но несколько более выраженный: R^2=0,119. Самым главным при этом является то, что во всех случаях не прослеживается зависимости какого-либо другого характера (квадратической, кубической и т.д.), что привело бы к невозможности использования соответствующей переменной в линейном регрессионном анализе. Степень же линейной взаимосвязи может быть результатом конфаундинг эффекта, суть которого заключается в том, что если какой-либо показатель (конфаундер) связан с одной из независимых переменных и сам по себе влияет на зависимую переменную и его влияние не оценивается одновременно со связанным с нимпредиктором, то связь последнего с переменную отклика может быть выявлена неточно [3]. Иногда наличие таких переменных (конфаундеров), которые не учтены в ходе анализа, или в процессе исследования в целом, могут не только увеличивать или наоборот занижать степень истинной взаимосвязи между переменными, но и нивелировать ее или изменять ее направление. В этом заключается одно из преимуществ множественной регрессии над простой - возможность оценки независимого друг от друга влияния на переменную отклика каждой из переменных-предикторов, включенных в модель.

Кроме двухмерной, SPSS позволяет использовать трехмерную скаттерограмму (3-D scatterplot) для того, чтобы визуально оценить взаимосвязь между тремя переменными одновременно («3-D Scatter», рис. 1).

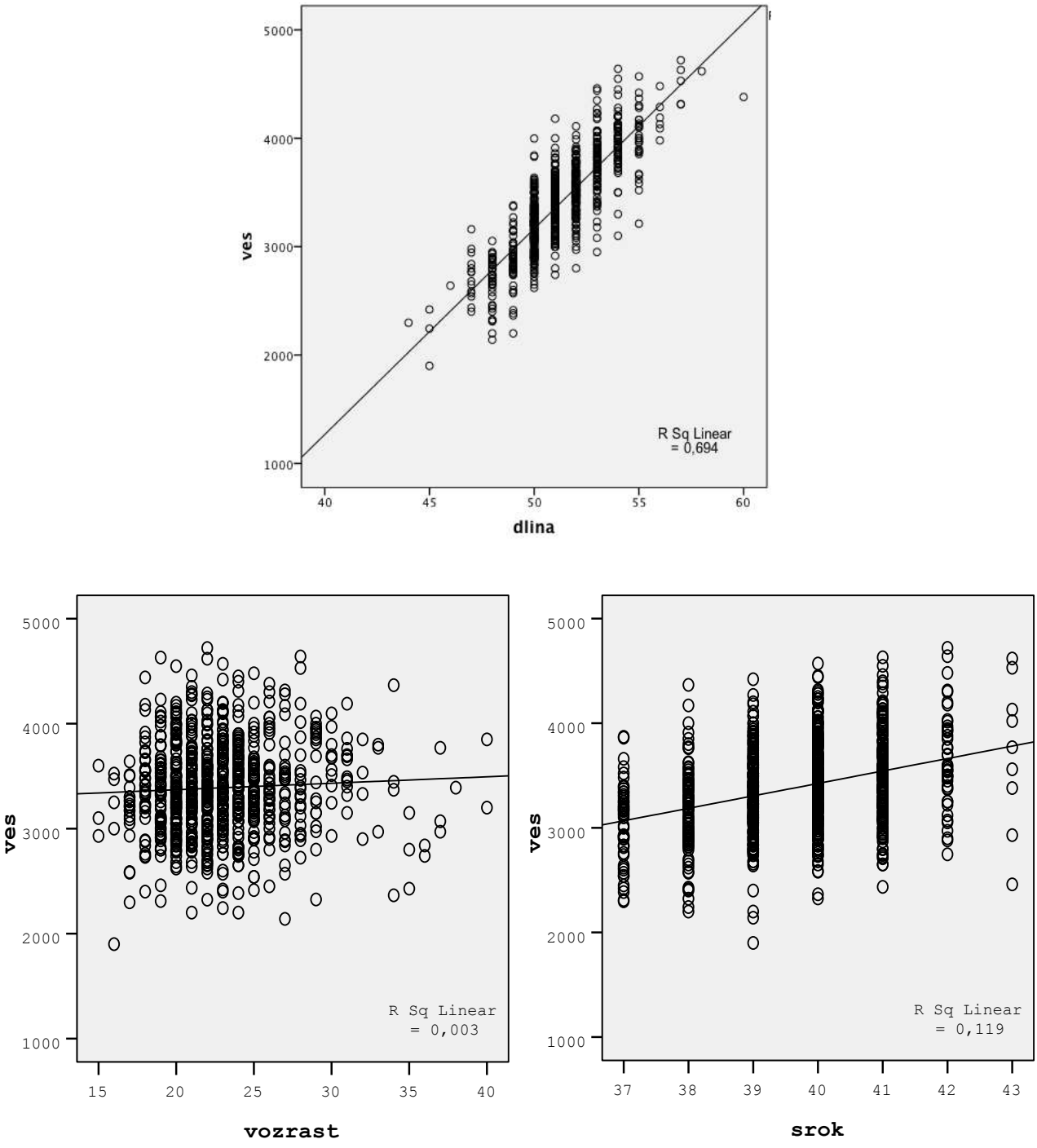

Рисунок 3. Скаттерограммы взаимосвязи между массой новорожденных и их длиной (вверху), между массой новорожденных и возрастом матери (слева внизу), между массой новорожденных и гестационным сроком (справа внизу), в г. Северодвинске.

Помимо наличия линейной взаимосвязи между переменными мы видим из скаттерограмм, что существует по несколько наблюдений с определенными значениями независимых переменных, но с различными значениями зависимых переменных. Другими словами, существует определенный разброс, или рассеяние значений зависимой переменной «ves» в зависимости от определенного значения переменных-предикторов. Мерой этого разброса в множественной линейной регрессионной модели, также как и в простой линейной регрессии, является сумма квадратов (или сумма вариаций). Разделение суммы квадратов подобно тому же в простой линейной регрессии и соответствует формуле [1, 3]:

$$SS_t = SS_m + SS_r$$

Общая сумма квадратов (total sum of squares SS_t), отражающая общую вариабельность переменной отклика, – это разница между фактическими значениями зависимой переменной и ее средним значением. Сумма квадратов модели (model sum of squares, SS_m), или та вариабельность зависимой переменной, которая объясняется моделью, рассчитывается как разность между значениями зависимой переменной, предсказываемыми моделью, и средним значением. Остаточная сумма квадратов (residual sum of squares, SS_r), отражающая вариабельность зависимой переменной, которая не может быть объяснена моделью (мера неточности построенной модели) равна разнице между фактическими значениями переменной и значениями, предсказанными моделью.

Таким образом, доля общей вариабельности переменной отклика, которую может объяснить регрессионная модель с несколькими переменными-предикторами, выражается в виде коэффициента детерминации (R^2), который рассчитывается и интерпретируется также как и в простой регрессии [1, 3]:

$$R^2 = \frac{SS_m}{SS_t}$$

SPSS рассчитывает значения все этих показателей для множественных моделей автоматически.

Для того, чтобы выполнить множественный регрессионный анализ необходимо проверить соблюдение всех необходимых условий [3-5], к которым относятся:

1. Независимость наблюдений;
2. Непрерывная зависимая переменная;
3. Линейная зависимость между переменной отклика и каждой независимой переменной;
4. Дисперсия каждой из независимых переменных >0;
5. Отсутствие мультиколлинеарности, т.е. ситуаций, когда независимые переменные сильно коррелируют между собой (r >0,9);
6. Независимость остатков;
7. Нормальное распределение остатков с M=0;
8. Гомоскедастичность, или одинаковое рассеяние остатков при любом предсказанном значении зависимой переменной.

Наблюдения в Северодвинском исследовании являются независимыми. Это определяется дизайном исследования. Данное исследование не является исследованием типа «до-после», или исследованием с подбором пар и т.д. Кроме того большое значение имеет правильность создания выборки. В идеале, каждый член популяции должен иметь одинаковую вероятность быть включенным в исследование. Это называется случайным отбором. В результате такого отбора формируется репрезентативная выборка, т.е. выборка, достаточно точно отражающая основные характеристики исходной популяции. Зависимая переменная («ves») измеряется с помощью интервальной шкалы, т.е. непрерывная. Мы выяснили, что условие линейной зависимости между переменной отклика и каждой независимой переменной соблюдается.

Для проверки оставшихся пяти условий нам потребуется SPSS, причем это можно сделать, непосредственно выполнив множественный регрессионный анализ.

Давайте построим множественную регрессионную модель для нашего примера. Она будет иметь следующий вид:

$$«ves»_i = (b_0 + b_1 \cdot «dlina»_i + b_2 \cdot «srok»_i + b_3 \cdot «vozrast»_i) + \varepsilon_i.$$

После того, как SPSS автоматически рассчитает значения всех трех b-коэффициентов и уровень их статистической значимости, мы с определенной степенью неточности (ε_i) сможем предсказывать значения массы тела ($«ves»_i$) для каждого конкретного ребенка (i), не только на основании его длины ($«dlina»_i$), но также в зависимости от гестационного срока ($«srok»_i$) и возраста матери ($«vozrast»_i$).

Для выполнения множественного линейного регрессионного анализа в SPSS необходимо открыть то же самое диалоговое окно «Linear Regression», что и при выполнении простого линейного регрессионного анализа. Для этого на панели инструментов необходимо выбрать меню «Analyze», в нем – раздел «Regression», затем «Linear». В открывшемся диалоговом окне (рис. 4) нужно выбрать зависимую переменную

«ves», кликнув на нее мышью, и перенести ее в поле «Dependent», путем нажатия на стрелку рядом с этим полем. Все независимые переменные, в нашем случае их три: «dlina», «srok» и «vozrast», должны быть перенесены в поле «Independent». Но при выполнении множественного анализа, когда мы имеем несколько независимых переменных, а не одну, большое значение имеет какие и сколько переменных выбрать и каким способом вводить их в модель. В идеальном варианте выбор переменных и введение их в модель должны быть основаны на результатах ранее проведенных исследований, на каких-либо теориях или гипотезах. Не следует отбирать сотни случайных предикторов, вводить их в модель и смотреть, что из этого получиться. Так как в реальности подавляющее большинство предикторов в той или иной степени влияет друг на друга и коррелирует между собой, результаты всякий раз могут быть различными.

Программа SPSS предлагает несколько способов ввода независимых переменных в модель. Каждый из них имеет свои особенности и предпочтителен в тех или иных ситуациях.

Метод форсированного, или одновременного ввода (Forced entry или Enter) – это метод, при котором все независимые переменные вводятся в модель одновременно (одним блоком). Этот метод основан на какой-то определенной теории или гипотезе для включения предикторов в модель, при этом исследователь не может определить порядок введения переменных в модель. Этот метод введения переменных предпочтителен в ситуациях, когда основная цель построения регрессионной модели не предсказание или прогнозирование, а оценка и сравнение степени независимого друг от друга влияния нескольких переменных-предикторов на переменную отклика, т.к. этот способ позволяет устранить конфаундинг эффект со стороны включенных в модель переменных.

При использовании метода иерархического, или блочного ввода (Hierarchical или Blockwise Entry), предикторы также отбираются на основании каких-то имеющихся знаний, однако здесь исследователь сам решает в каком порядке вводить их в модель. Как правило, известные предикторы, влияние которых на переменную отклика уже было доказано в других исследованиях, вводятся в модель первыми (первым блоком), т.е. в порядке важности или степени их непосредственного влияния на зависимую переменную. После того, как эти переменные уже введены, исследователь может добавить в модель новые, основываясь на определенной гипотезе. Причем новые переменные могут быть добавлены в модель общим блоком, или несколькими блоками (иерархически), или один за другим (пошагово). Это также зависит от цели и задач исследования. Например, если предполагается, что влияние ряда новых предикторов на переменную отклика имеет наибольшую важность, то они могут быть введены первыми и т.д.

Кроме перечисленных двух, SPSS предлагает ряд пошаговых способов введения переменных. Это метод последовательного ввода (Forward), метод пошагового ввода (Stepwise) и метод последовательного исключения (Backward). Особенностью всех пошаговых способов является то, что самостоятельно исследователь только выбирает ряд интересующих его предикторов, а порядок, в котором они будут введены в модель, определяется самой программой исключительно на основании математических критериев. Так, при использовании метода последовательного ввода (Forward) начальная модель содержит только константу (b_0). Затем программа выбирает из предложенных ей предикторов тот, который в наибольшей степени коррелирует с зависимой переменной (следовательно, лучше, чем все остальные, предсказывает ее). Если включение этой переменной статистически значимо улучшает предсказательную способность модели, тогда программа оставляет ее в качестве зависимой переменной и ищет следующий предиктор. Следующим отбирается тот предиктор, который объясняет большую часть оставшейся вариабельности зависимой переменной, т.е. той, которая не может быть объяснена предыдущей моделью с одним предиктором (semi-partial correlation). Если после включения этого предиктора в модель ее способность предсказывать зависимую переменную улучшается статистически значимо (R^2 модели увеличивается статистически значимо по сравнению с R^2 предыдущей модели), то второй предиктор остается в модели, а программа ищет

следующий по аналогичной схеме. Когда включение очередного предиктора не приводит к значимому улучшению предсказательной способности модели, он из нее удаляется, и поиск новых предикторов прекращается.

Метод пошагового ввода (Stepwise) в SPSS аналогичен предыдущему, с той лишь разницей, что при включении каждого нового предиктора в модель из нее удаляется наименее сильный предиктор и проверяется, приводит ли это к статистически значимому уменьшению предсказательной способности. Если да, то он возвращается в модель, если нет, то удаляется из нее. Таким образом, регрессионное уравнение постоянно подвергается переоценке для того, чтобы проверить, могут ли какие-то из включенных ранее предикторов быть удалены без ущерба предсказательной способности модели. Наличие такой возможности связано с тем, что значительная часть вариабельности зависимой переменной, которая объяснялась ранее включенными предикторами, на самом деле была обусловлена другими переменными (конфаундинг эффект), и после их включения в модель сильно уменьшилась.

Метод последовательного исключения (Backward) – это противоположность метода последовательного ввода. Другими словами, суть его в том же, но программа начинает с введения всех потенциальных предикторов в модель и рассчитывает долю, вносимую каждым из них в предсказательную способность модели (t-тест для каждого предиктора). Если предиктор не вносит статистически значимого вклада в способность модели предсказывать зависимую переменную, он удаляется из модели, и последняя переоценивается. Также подвергаются повторной оценке оставшиеся в модели предикторы.

Пошаговые методы предпочтительнее использовать в тех случаях, когда цель построения множественной регрессионной модели – предсказание значения независимой переменной по значениям нескольких переменных-предикторов, т.е. когда необходимо выбрать по возможности наименьшее количество предикторов, которые обеспечат максимальную точность предсказания. Из всех пошаговых методов метод последовательного исключения предпочтителен, т.к. он несет наименьший риск ошибки второго типа (наименьшую вероятность исключить из модели предиктор, который на самом деле оказывает влияние на зависимую переменную). С чем это связано? Существуют такие ситуации, когда предиктор оказывает статистически значимый эффект только в присутствии другой переменной в качестве предиктора (suppressor effect) [3]. В этом случае при использовании метода последовательного ввода существует большая вероятность удалить эти переменные из модели, тогда как при использовании метода последовательного исключения эти предикторы в ней останутся.

В целом, при проведении множественного регрессионного анализа рекомендуется избегать пошаговых методов за исключением случаев эксплораторного анализа, т.к. при этом множество методологических решений, которые должны приниматься исследователем на основании уже доказанных данных, теорий и тестируемых гипотез, принимаются программой исключительно на основании математических критериев. Однако то, что программа посчитала неважным или незначительным с математической точки зрения, может иметь огромное теоретическое значение.

Как все-таки вводить переменные в модель? При построении модели всегда опирайтесь на данные, полученные в предыдущих исследованиях. Включайте в модель переменные в порядке их важности, блоками или по-одному. После первоначального анализа повторите регрессию, исключив те переменные, которые не имели значимого влияния в первый раз. Затем, исходя из теоретической важности и статистической значимости, решите, какие переменные должны быть включены в модель. Не нужно стремиться включать максимально возможное число независимых переменных из соображений статистической мощности. Как правило, чем меньше предикторов, тем лучше. Кроме того, включайте только те переменные, для которых имеется хорошее теоретическое обоснование. И не забывайте о достаточном объеме выборки: должно быть не менее 15-20 наблюдений на каждый предиктор [3, 4, 10].

При проведении простого регрессионного анализа уже было выяснено, что «dlina» оказывает влияние на переменную отклика «ves» [1].

Две другие переменные, «srok» и «vozrast», согласно нашей гипотезе также могут влиять на массу тела при рождении, хотя и в меньшей степени. Следовательно, используя иерархический метод ввода независимых переменных в модель, первым блоком введем длину, как наиболее важную переменную-предиктор, а вторым блоком две другие. Для этого в уже открытом окне «Linear Regression» с уже обозначенной зависимой переменной «ves» (рис. 4), выделяем переменную «dlina» и переносим ее из левого поля в поле «Independent(s)» слева, кликнув на соответствующую стрелку. При этом становится активной кнопка «Next». Нажав на последнюю, мы открываем окно «Independent(s)» для второго блока (Block 2 of 2), в которое мы переносим две оставшиеся независимые переменные – «srok» и «vozrast». Нажимая на кнопки «Previous» и «Next» можно возвращаться к предыдущему или переходить в следующий блок независимых переменных, если это необходимо. Выпадающее окно меню под названием «Method» необходимо, если Вы хотите использовать методы пошагового ввода. По умолчанию в нем установлено Enter, т.е. метод форсированного ввода, который можно заменить на требующийся, открыв его и выбрав нужный нажатием.

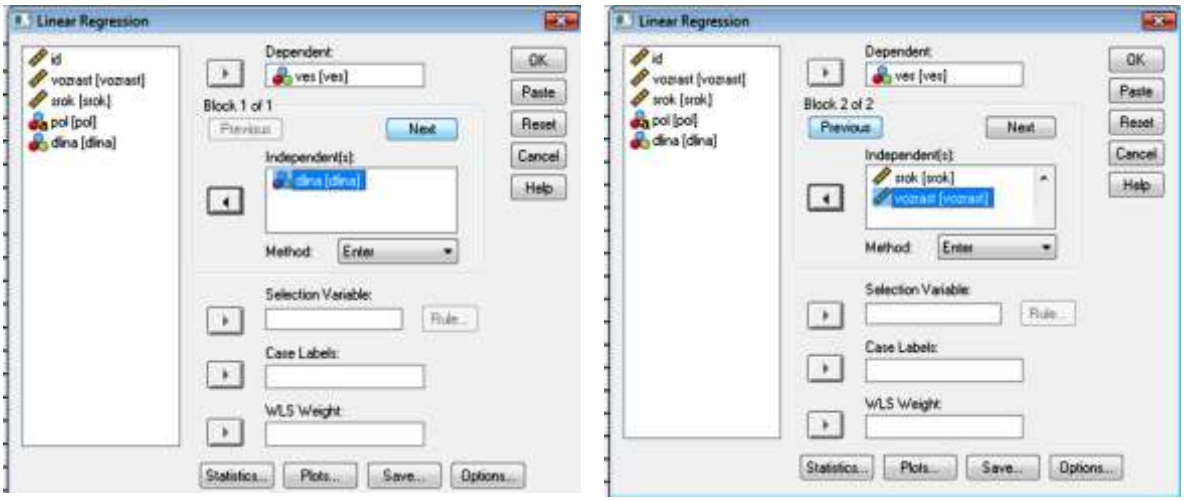

Рисунок 4. Окно «Linear Regression» после введения первого блока независимых переменных (слева) и после второго блока (справа).

После того, как определены зависимая и независимые переменные, а также способ их ввода в модель, в меню «Statistics» (рис. 5) необходимо отметить следующие пункты: Estimates (для оценки параметров путем проверки гипотезы о равенстве параметров «0»), Model fit (для расчета R, R^2, Adjusted или скорректированного R^2, критерия F для модели и его статистической значимости), Confidence intervals (для оценки доверительных интервалов параметров модели), Descriptives (показывает средние значения всех переменных и Correlations (корреляционный матрикс), Collinearity diagnostics (для расчета VIF и Tolerance и др. способов оценки мультиколлинеарности), критерий Durbin-Watson (критерий для оценки независимости остатков), Casewise diagnostics (для оценки выскакивающих наблюдений; по умолчанию – это те наблюдения, остатки которых превышают 3 стандартных отклонения; данное значение желательно изменить на 2) и R squared change (для оценки изменения предсказательной способности модели при изменении числа предикторов или блоков). В меню «Plots» (рис. 6) необходимо отметить Histogram и Normal probability plot (для визуальной оценки нормальности распреде-ления остатков), Produce all partial plots (для проверки наличия линейной зависимости между переменной отклика и каждой из переменных предикторов), а также перенести ZRESID в Y-окно, а ZPRED – в X-окно (построение скаттерограммы зависимости стандартизованных остатков от стандарти-зованных предсказанных значений для проверки условия гомоскедастичности). В меню «Save» (рис. 7) отмечаем Standardized residuals, Cook's distance, Leverage values, Standardized DfBeta(s), Covariance ratio, что необходимо для диагностики модели (подробнее будет рассмотрена далее). Выполнение множественного регрессионного анализа запускаем нажатием на кнопку «ОК».

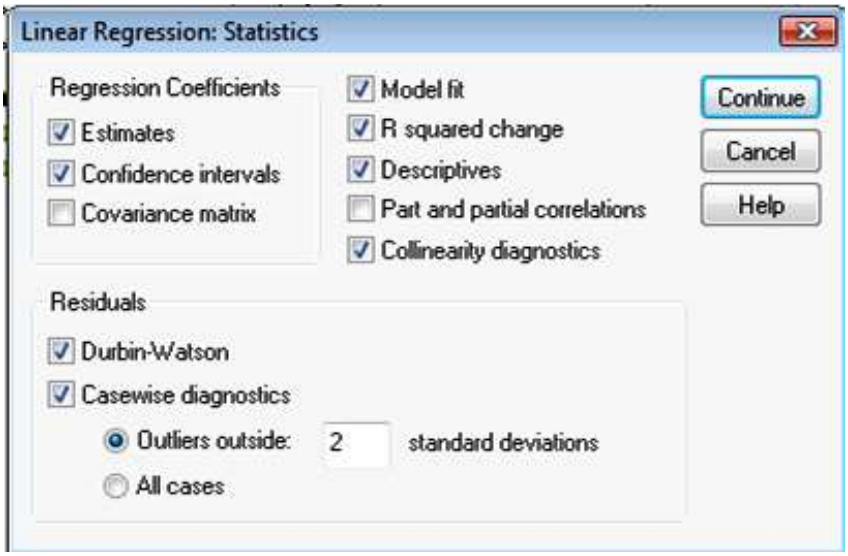

Рисунок 5. Диалоговое окно «Linear Regression: Statistics».

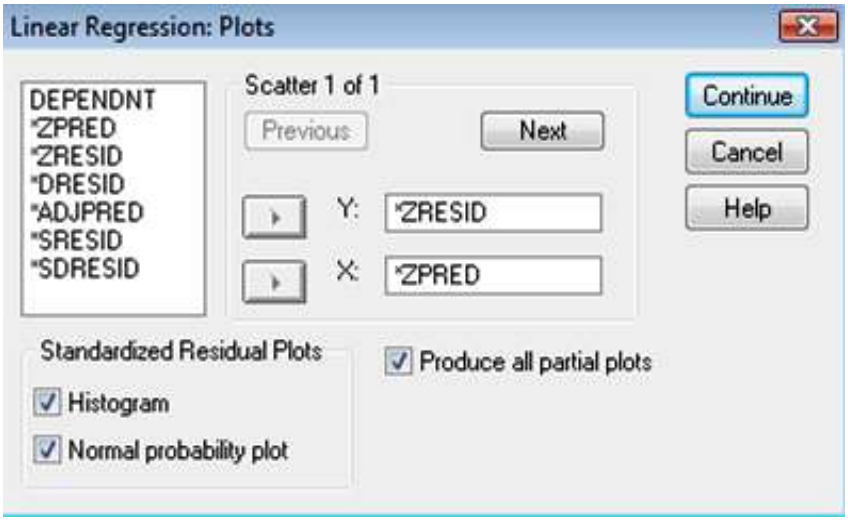

Рисунок 6. Диалоговое окно «Linear Regression: Plots».

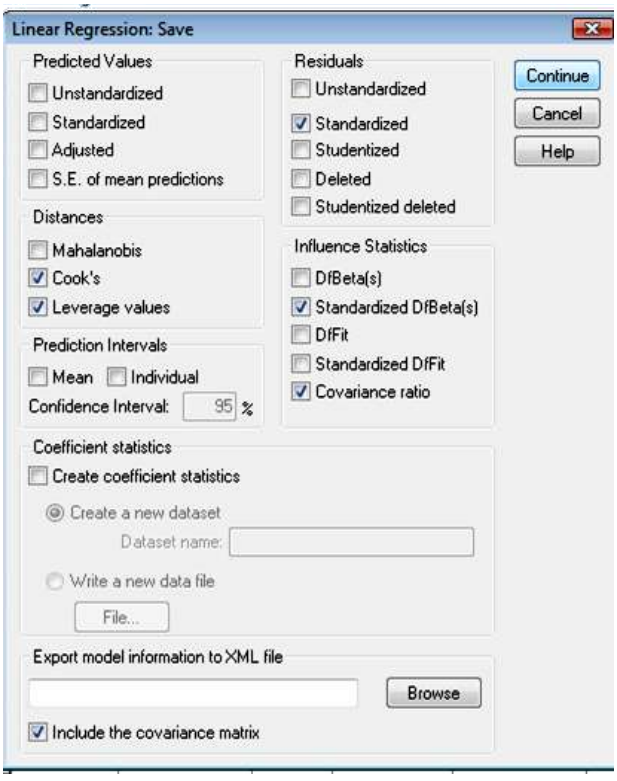

Рисунок 7. Диалоговое окно «Linear Regression: Save».

Оценка результатов анализа. Первые две таблицы аналогичны таковым при проведении простой линейной регрессии [1]. Первая таблица (рис. 8) показывает данные описательной статистики (Descriptive Statistics) для всех четырех переменных, включенных в модель. По значениям стандартных отклонений (Standard deviation) для независимых переменных, представленных в этой таблице, мы можем косвенно судить о значениях их дисперсий. Дисперсия равна квадрату стандартного отклонения. Следовательно, если стандартные отклонения для всех независимых переменных отличаются от 0 (1,913 – для «dlina», 1,266 – для «srok» и 3,655 – для «vozrast»), то и их дисперсии будут отличаться от нуля. Это является четвертым условием множественного регрессионного анализа. Существует другой способ проверки этого условия: открыть окно «Descriptives» (Analyze, Descriptive statistics, Descriptives), переместить все три независимые переменные из левого поля в правое «Variable(s)», отметить в меню «Options» Variance и затем запустить выполнение анализа нажатием на кнопку «OK». Дисперсии (Variances) всех трех переменных отличаются от нуля.

Вторая таблица (рис. 9) представляет собой корреляционный матрикс (Correlations), показывающий коэффициенты корреляции Пирсона для оценки линейной связи переменных друг с другом и уровень ее статистической значимости, который должен быть умножен на 2, т.к. принято представлять двусторонние тесты (2-tailed), а не односторонние (1-tailed). По этой таблице мы можем проверить пятое условие отсутствия мультиколлинеарности: для нашего примера r – коэффициенты корреляции для независимых переменных «dlina» и «srok», «dlina» и «vozrast», «vozrast» и «srok» - равны 0,327, 0,068 и 0,052 соответственно. Все они меньше 0,9.

Третья небольшая таблица «Variables Entered/Removed», изображенная на рис. 10, показывает, какие переменные вводились (entered) в модель при каждом следующем шаге (если использовались иерархический или пошаговый способы ввода независимых переменных), или исключались (removed) из нее, что бывает в случаях несоблюдения некоторых условий включения переменных. Таким образом, в нашем примере в первую модель была введена переменная «dlina», а во вторую модель были добавлены переменные «vosrast» и «srok», при этом ни одна переменная из модели не удалялась. Сноска «a» в обоих случаях (и в первой, и во второй модели) говорит о том, что все переменные, которые исследователь предполагал ввести в модель на соответствующем этапе, были в нее включены, т.е. удовлетворяли условию, которое установлено в SPSS по умолчанию: probability of F >0,05 (критерий для включения переменной), но <0,01 (критерий для исключения переменной). Оба этих крайних значения можно изменить, зайдя в меню «Options» в правом нижнем углу окна «Linear Regression» (рис. 4), но обычно этого делать не рекомендуется. Сноска «b» говорит о том, что зависимая переменная в модели «ves».

Таблица «Model Summary» на рис. 11 при множественной линейной регрессией также представляет общую информацию о модели, причем здесь представлена информация сначала о первой с одной независимой переменной «dlina», которую мы вводили первым блоком, а затем о второй модели с тремя предикторами («dlina» плюс переменные из второго блока – «vozras» и «srok»). В этой таблице мы можем видеть, что значение R^2, для второй модели больше, чем для первой. Следовательно, наша новая модель с двумя дополнительными предикторами объясняет уже не 69,4%, а 70,0% вариабельности массы новорожденного. R^2 стал больше всего на 0,6% (R Square change для модели 2 равен 0,006, рис. 12), но даже такое небольшое увеличение предсказательной способности модели является статистически значимым (Sig. F Change для модели 2 < 0,001). R Square change для первой модели, равное значению самого R square (0,694, или 69,4%), показывает на сколько изменяется предсказательная способность при использовании регрессионной модели по сравнению с использованием среднего арифметического в качестве предсказательной модели. Обратим внимание, что коэффициенты детерминации простого регрессионного анализа уже были автоматически указаны на скаттерограммах (рис. 3). И это изменение также статистически значимо (Sig. F Change для модели 1 < 0,001).

Другими словами, простая регрессионная модель с одним предиктором «dlina» предсказывает значение зависимой переменной «ves» статистически значимо лучше, чем просто использование средней - 3388,2 (рис. 8), а множественная регрессионная модель с предикторами «dlina», «srok» и «vozrast» - статистически значимо лучше, чем простая модель 1.

В этой же таблице (рис. 11) представлен показатель adjusted R Square (скорректированный R^2), значение которого совпадает со значением нескорректированного, или грубого R^2 для первой модели и несколько ниже последнего для второй модели. Скорректированное значение показывает сокращение предсказательной мощности. Грубое значение показателя говорит нам, какая доля вариабельности Y может быть объяснена регрессионной моделью, построенной на данных нашей выборки. Скорректированное значение говорит о том, какую долю вариабельности Y объясняла бы эта модель, если бы она была построена на данных всей популяции, из которой была извлечена выборка [3-5]. Вы можете рассчитать значение скорректированного R^2 самостоятельно, используя формулу Стейна (Stein`s) [3]:

$$\text{adjusted } R^2 = 1-[((n-1)/(n-k-1)) \cdot ((n-2)/(n-k-2)) \cdot ((n+1)/n)] \cdot (1- R^2),$$

где n – количество наблюдений, k – количество предикторов в модели, R^2 – грубое значение коэффициента детерминации.

SPSS рассчитывает значение скорректированного R^2, используя другую формулу (Wherry`s equation) [11]:

$$\text{adjusted } R^2 = [1-(1- R^2)\cdot(n-1)/(n-k-1)]$$

Формула Wherry менее благоприятна для кроссвалидации, т.к. она не может предсказать, насколько хорошо модель может предсказывать значения зависимой переменной для других выборок из той же популяции.

Значение критерия Дарбина-Уотсона (Durbin-Watson), указанное в этой же таблице, равно 2,039, что говорит о том, что условие независимости остатков (№6) соблюдается. В идеальном варианте значение критерия должно быть равно 2, но допускаются значения от 1 до 3.

Две следующие таблицы – «ANOVA» (рис. 12), отражающая результаты применения критерия F для проверки значимости модели, и таблица «Coefficients» (рис. 13), представляющая информацию о коэффициентах регрессии, их доверительных интервалах и статистической значимости, – также представляют информацию сначала для первой модели с первым блоком предикторов, а затем для второй модели с двумя блоками. Таким образом, информация о первой модели (Model 1) аналогична информации, полученную при проведении простой линейной регрессии [1] и необходима нам, чтобы оценить, насколько сильное влияние оказало включение в модель новых предикторов. Информация же о второй модели, представляющей наибольший интерес, содержится в таблицах под цифрой 2 (Model 2).

Так, в таблице на рис. 12 под цифрой 2 представлены значения сумм квадратов (суммы квадратов модели, суммы квадратов остатков и общей суммы) и соответствующие степени свободы, необходимые для расчета средних квадратов – Mean Square (отношение суммы квадратов к соответствующему количеству степеней свободы). Значение коэффициента F, равное 673,563 получено путем деления среднего квадрата для модели, отражающего долю вариабельности зависимой переменной, которую можно объяснить нашей моделью, на средний квадрат остатков, отражающий долю вариабельности зависимой переменной, которая не может быть объяснена нашей моделью. Другими словами, если значение коэффициента F больше 1, то доля вариабельности, объясняемая моделью, больше той, которая не может быть ею объяснена, и чем больше это значение по сравнению с единицей, тем лучше наша модель. Если полученное значение F превышает критическое, которое определяется для соответствующих степеней свободы (в нашем случае количество степеней свободы (degrees of freedom, df = 865 и 3, рис. 12) и которое можно найти в

соответствующих таблицах в справочниках, это означает, что регрессионная модель статистически значима при соответствующем уровне значимости (чаще всего это 5% уровень). SPSS автоматически сравнивает фактическое значение F с критическим и представляет абсолютный уровень его статистической значимости – Sig. (в нашем случае p<0,001).

Таблица регрессионных коэффициентов (рис. 13) представляет абсолютные значения константы (b_0) и регрессионных коэффициентов для каждого предиктора (Unstandardized Coefficient, B), а также уровень их статистической значимости (Sig.), полученный в результате проверки нулевой гипотезы о равенстве каждого из них нулю с помощью критерия Стьюдента (t-тест). Таким образом, переменные «dlina» и «srok» оказывают статистически значимое влияние на массу тела новорожденного, т.е. нулевая гипотеза о равенстве их регрессионных коэффициентов нулю отвергается при 5% уровне значимости, в то время как возраст матери не оказывает статистически значимого влияния. Уровень Sig. для этой переменной равен 0,684. Это говорит о том, что на самом деле с вероятностью 68,4% коэффициент регрессии для этой переменной равен нулю, следовательно, при 5% уровне значимости мы принимаем нулевую гипотезу.

Значения константы и коэффициентов регрессии для каждой переменной интерпретируются таким же образом, как и в простой линейной регрессии. Значение константы показывает значение переменной отклика, если значение предиктора будет равно нулю. Значение коэффициента регрессии показывает, насколько увеличится значение переменной отклика, если значение предиктора увеличится на единицу. На примере модели 2 (рис. 13) видим, что масса тела новорожденных увеличится на 183,8 см при увеличении длины при рождении на 1 см.

Следует отметить, что значение коэффициента регрессии для переменной «dlina» уменьшилось с 189,83 (в первой модели) до 183,85 (во второй модели). Такое часто бывает с коэффициентами при включении новых переменных в модель. Это обусловлено тем, что частично влияние длины тела на массу объяснялось влиянием гестационного срока (конфаундер) и, возможно, возрастом матери, и после включения последних в многофакторный анализ их влияние было устранено.

Кроме того, в связи с тем, что регрессионные коэффициенты измеряются в тех же единицах, что и сами переменные, для сравнения степени влияния каждого из них на переменную отклика они не могут быть использованы (размер коэффициента зависит от единиц измерения). Для этой цели можно использовать стандартизованные коэффициенты (Standardized Coeficients), единицы измерения которых одинаковы – количество стандартных отклонений. Стандартизованные коэффициенты показывают, на сколько стандартных отклонений увеличивается переменная отклика при увеличении предиктора на одно стандартное отклонение. Так, мы видим, что влияние длины тела (станд. b = 0,807) на массу новорожденного практически в 10 раз превышает степень влияния гестационного срока (станд. b = 0,082).

Помимо уровня статистической значимости коэффициентов таблица представляет их 95% доверительные интервалы (95% Confidence Interval for B). Они говорят о том, в каких пределах с 95% вероятностью находится популяционное значение коэффициента. Чем шире интервал, тем менее точно значение коэффициента b отражает его значение для генеральной совокупности, которой соответствует изучаемая выборка, и наоборот. В тоже время, по тому включает ли интервал 0 или нет, мы можем также судить о статистической значимости соответствующего коэффициента. Если с 95% вероятностью коэффициент не может быть равен нулю (интервал не включает 0), то он статистически значим и наоборот. Так, для переменной «vosrast» 95% доверительный интервал от -5,275 до 3,463, т.е. включает 0, что, также как и значение p = 0,684, свидетельствует об отсутствии значимости при 5% уровне. В данном примере коэффициент регрессии переменной «dlina» для генеральной совокупности будет с 95% надежностью находиться в пределах от 175,0 до 192,6 см.

Таблица коэффициентов кроме всего перечисленного представляет информацию о значениях VIF и толерантности (Tolerance), которые необходимы для проверки пятого условия отсутствия мультиколлинеарности. Для соблюдения этого условия необходимо,

чтобы все значения VIF были менее 10, а Tolerance, которое обратно пропорционально VIF (Tolerance = 1/VIF), – более 0,1 [3-5]. Таким образом, мы еще раз подтвердили, что условие отсутствия мультиколлинеарности в нашем примере соблюдается (рис. 13).

В таблице на рис. 14 представлены те наблюдения, остатки которых выходят за пределы трех стандартных отклонений (если Вы не меняли установленное в SPSS по умолчанию значение 3 в графе Casewise diagnostics, меню «Statistics») или двух, (если это значение было изменено на 2, как в нашем случае, и как рекомендуется делать). Такие случаи называют «выскакивающими» (outliers). Число случаев, остатки которых выходят за пределы двух или трех стандартных отклонений, не должно превышать 5% или 1% от общего объема выборки, соответственно. В противном случае наша модель плохо соответствует имеющимся данным или обладает низкой предсказательной точностью. Кроме того, необходимо обращать внимание на знаки этих остатков: число положительных значений должно быть примерно равно числу отрицательных. Если остатки почти всех «выскакивающих» случаев имеют одинаковый знак, то эти случаи требуют отдельного рассмотрения, т.к. представляют собой в своем роде обособленный класс. В нашем исследовании имеется 6 случаев, остатки которых выходят за пределы 2 стандартных отклонений, что составляет 0,7%, что значительно меньше 5%, причем четыре остатка имеют знак «-» и два – знак «+». Таким образом, наша модель вполне соответствует перечисленным требованиям.

Существуют и другие способы оценить наличие «выскакивающих» случаев. Один из них визуальный: мы можем увидеть их на скаттерограмме, хотя в случае с множественной регрессионной моделью это может представлять некоторые трудности. Второй способ: посмотреть у скольких наблюдений стандартизованные остатки, которые были сохранены в базе при выполнении анализа под названием «ZRE», выходят за пределы -/+ 1,96 (должно быть не более 5% от выборки), за пределы -/+ 2,58 (не более 1%) или за пределы -/+ 3,29 (не более 0,1%).

Помимо «выскакивающих» случаев, существуют так называемые случаи, оказывающие сильное влияние на модель (influential cases) [3]. Это такие случаи, которые тем или иным образом отличаются от общей массы наблюдений и оказывают сильное влияние на параметры модели: значение константы, регрессионных коэффициентов, их статистическую значимость и т.д. Другими словами, они смещают регрессионную прямую на себя, в результате чего размер их остатков не отличается значительно от остальных. Если такой случай или случаи исключить из анализа и построить новую модель без них, то параметры ее значимо изменятся. Выявление таких случаев необходимо для того, чтобы определить, насколько устойчива модель, т.е. не смещена ли она под влиянием отдельных наблюдений. Для этой цели SPSS для каждого отдельного наблюдения рассчитывает и сохраняет несколько коэффициентов, позволяющих оценить степень его влияния на модель. Это Cook's distance или дистанция Кука (сохраняется в базе под названием «COO»), Leverage или «рычаг» (в базе – «LEV»), DfBeta и стандартизованный DfBeta для константы и каждого из предикторов («SDB»).

Cook's distance позволяет оценить степень влияния каждого случая на модель в целом. Принято, что случаи, для которых этот коэффициент имеет значение больше 1, являются «оказывающими сильное влияние» и требуют отдельного рассмотрения [12].

Второй мерой является Leverage. Этот показатель также оценивает влияние каждого фактического значения переменной отклика на предсказываемые значения, но является несколько более чувствительным, чем дистанция Кука. Среднее значение этого показателя рассчитывается по формуле:

$$(k + 1)/n,$$

где k – количество предикторов в модели, а n – число наблюдений.

Leverage для каждого случая может принимать значения от 0 (случай не оказывает никакого влияния на модель вообще) до 1 (случай имеет абсолютное влияние на предсказание). В идеале, если никакие наблюдения не оказывают сильного влияния на модель, значения этого показателя должны быть близки к его среднему значению (см. формулу выше). Следует обращать внимание

на случаи, имеющие значение Leverage, в два [13] или в три раза [14] превышающее среднее значение, т.к. они могут оказывать сильное влияние на модель. Однако необходимо отметить, что не всегда случаи с большим значением Leverage влияют на регрессионные коэффициенты, т.к. при его расчете используются единицы зависимой переменной, а не переменных-предикторов.

Если выполнить регрессионный анализ, исключив какой-либо случай, оказывающий влияние на модель, регрессионные коэффициенты модели изменятся. По этой разнице можно судить о степени влияния отдельного случая на модель и ее предсказательную способность. Разница между параметром модели, полученным при выполнении регрессионного анализа на всей выборке, и этим же параметром, полученным после исключения из анализа отдельного случая, рассчитывается в SPSS и сохраняется в базе под названием DFBeta для каждого случая и для каждого регрессионного коэффициента, включая константу (b_0). Таким образом, оценив значения DFBetas, мы также можем выявить случаи, оказывающие влияние на регрессионную модель. Но здесь опять же необходимо учитывать: размер показателей будет зависеть от единиц измерения каждой из независимых переменных, и поэтому сложно определить универсальное критическое значение показателя, выше которого случаи будут считаться «оказывающими сильное влияние на модель». В связи с этим рекомендуется использовать стандартизованные показатели (standardized DFBeta) и их пороговое значение, равное, согласно различным источникам, >|1| или >|2| [3, 14].

Последний показатель для оценки влияния отдельных случаев на предсказательную способность модели – это CVR (covariance ratio). Он показывает, как влияет каждый отдельный случай на дисперсию (точность регрессионных коэффициентов). В идеальном варианте значение показателя должно быть близко к 1. Случаи, для которых CVR > 1 + [3(k+1)/n], где k – количество предикторов в модели, n – объем выборки, уменьшают дисперсию регрессионных коэффициентов, т.е. их удаление ухудшит точность оценки параметров. Удаление же случаев, для которых CVR < 1 - [3(k+1)/n], приводит к улучшению точности оценки параметров модели [3].

Таблица «Residuals Statistics» на рис. 15 частично представляет информацию, необходимую для выявления случаев, оказывающих сильное влияние на модель. В ней представлены минимальные (Minimum) и максимальные (Maximum) значения ряда показателей, их средние значения (Mean) со стандартным отклонением (Std. Deviation) и число наблюдений (N), на основании которых они рассчитывались. Нас в первую очередь интересуют такие показатели, как Cook's distance и Leverage, а именно их максимальные значения. Так, значения дистанции Кука для всех наблюдений в нашем примере находятся в пределах от 0,000 до 0,032 (все значительно меньше единицы), что говорит об отсутствии случаев, оказывающих сильное влияние на модель. Об этом же свидетельствует и крайнее значение Leverage. Оно равно 0,054, т.е. не превышает 3,004, рассчитанное по формуле: 3 + (3+1)/869. Таким образом, оба показателя говорят о том, что в нашем примере нет случаев, оказывающих сильное влияние.

Эти, а также ряд других описанных выше показателей, были рассчитаны при выполнении анализа и сохранены в базе для каждого наблюдения (рис. 16). Расположив наблюдения в возрастающем, а затем убывающем порядке для каждой из переменных, мы можем легко увидеть крайние значение каждого из показателей и сравнить их с пороговыми. Для этого необходимо нажать правой кнопкой мыши на название соответствующей переменной и в появившемся окне выбрать «Sort Ascending» для расположения случаев по возрастанию значений этой переменной, или «Sort Descending» для расположения их в убывающем порядке (рис. 16). После выполнения этого алгоритма для четырех стандартизованных DFBeta (для константы и трех регрессионных коэффициентов), мы видим, что их значения для всех случаев не превышают |1|. Все значения CVR (в базе COV) находятся в пределах от 0,950 до 1,038, т.е. достаточно близко к единице. Но существует ряд случаев, значение CVR для которых превышает 1 + [3(3+1)/869]=1,014 (n=39). Другими словами, их удаление ухудшает предсказательную точность модели.

Случаев со значением CVR < 1 - [3(3+1)/869]=-0,986 в нашем примере не было.

Помимо диагностики случаев, оказывающих сильное влияние на модель, таблица на рис. 16 несет описательную информацию о предсказываемых значениях зависимой переменной (Predicted Value), их стандартизованных (Std. Predicted Value) и скорректированных (Adjusted Predicted Value) значениях, остатках (Residual), стандартизованных остатках (Std. Residual) и т.д. При желании, SPSS также позволяет рассчитать и сохранить значения этих переменных для каждого наблюдения (рис. 7).

Descriptive Statistics

	Mean	Std. Deviation	N
ves	3388,20	435,806	869
dlina	51,18	1,913	869
srok	39,70	1,266	869
vozrast	22,83	3,655	869

Рисунок 8. Описательная статистика.

Correlations

		ves	dlina	srok	vozrast
Pearson Correlation	ves	1,000	,833	,346	,052
	dlina	,833	1,000	,327	,068
	srok	,346	,327	1,000	,052
	vozrast	,052	,068	,052	1,000
Sig. (1-tailed)	ves	.	,000	,000	,064
	dlina	,000	.	,000	,022
	srok	,000	,000	.	,065
	vozrast	,064	,022	,065	.
N	ves	869	869	869	869
	dlina	869	869	869	869
	srok	869	869	869	869
	vozrast	869	869	869	869

Рисунок 9. Коэффициенты корреляции для переменных «ves», «dlina», «srok» и «vozrast».

Variables Entered/Removed(b)

Model	Variables Entered	Variables Removed	Method
1	dlina(a)	.	Enter
2	vozrast, srok(a)	.	Enter

a All requested variables entered.

b Dependent Variable: ves

Рисунок 10. Переменные, включенные и удаленные из модели.

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	,833[a]	,694	,694	241,115	,694	1968,686	1	867	,000	
2	,837[b]	,700	,699	239,017	,006	8,644	2	865	,000	2,039

a. Predictors: (Constant), dlina

b. Predictors: (Constant), dlina, vozrast, srok

c. Dependent Variable: ves

Рисунок 11. Общие сведения о модели.

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1,1E+008	1	114452353,3	1968,686	,000[a]
	Residual	50404283	867	58136,428		
	Total	1,6E+008	868			
2	Regression	1,2E+008	3	38480000,39	673,563	,000[b]
	Residual	49416636	865	57129,058		
	Total	1,6E+008	868			

a. Predictors: (Constant), dlina

b. Predictors: (Constant), dlina, vozrast, srok

c. Dependent Variable: ves

Рисунок 12. Результат применения критерия F для определения значимости регрессионной модели.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	-6326,624	219,104		-28,875	,000	-6756,660	-5896,589		
	dlina	189,831	4,278	,833	44,370	,000	181,434	198,229	1,000	1,000
2	(Constant)	-7116,955	292,725		-24,313	,000	-7691,489	-6542,420		
	dlina	183,849	4,495	,807	40,898	,000	175,026	192,671	,890	1,123
	srok	28,143	6,784	,082	4,149	,000	14,829	41,458	,892	1,121
	vozrast	-,906	2,226	-,008	-,407	,684	-5,275	3,463	,994	1,006

a. Dependent Variable: ves

Рисунок 13. Таблица регрессионных коэффициентов.

Casewise Diagnostics(a)

Case Number	Std. Residual	ves	Predicted Value	Residual
436	3,512	4180	3340,61	839,390
709	-3,360	3212	4015,19	-803,189
717	-3,222	2200	2970,20	-770,196
743	-3,169	2950	3707,40	-757,401
789	3,531	3998	3154,04	843,955
796	-3,424	3100	3918,49	-818,488

a Dependent Variable: ves

Рисунок 14. Таблица «выскакивающих» случаев.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	1998,29	5044,29	3388,20	364,685	869
Std. Predicted Value	-3,811	4,541	,000	1,000	869
Standard Error of Predicted Value	8,473	43,484	15,363	5,195	869
Adjusted Predicted Value	1992,05	5062,26	3388,27	364,825	869
Residual	-818,488	843,955	,000	238,603	869
Std. Residual	-3,424	3,531	,000	,998	869
Stud. Residual	-3,432	3,534	,000	1,001	869
Deleted Residual	-822,279	845,626	-,075	239,826	869
Stud. Deleted Residual	-3,454	3,558	,000	1,002	869
Mahal. Distance	,092	27,730	2,997	3,097	869
Cook's Distance	,000	,054	,001	,003	869
Centered Leverage Value	,000	,032	,003	,004	869

a. Dependent Variable: ves

Рисунок 15. Таблица анализа остатков.

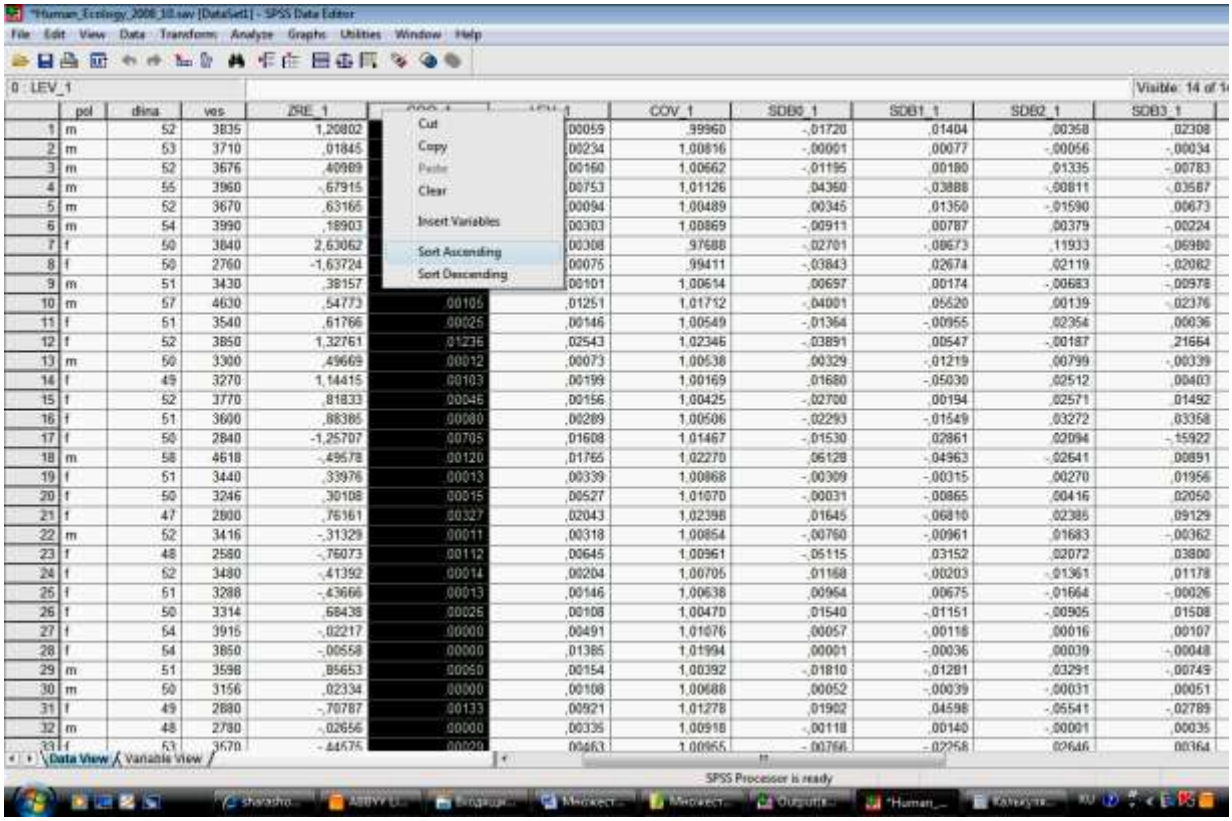

Рисунок 16. Вид в базе данных выбранных показателей, рассчитанных при проведении множественного линейного регрессионного анализа SPSS.

После ряда таблиц, которых мы только что рассмотрели, следует графическая информация, также необходимая, главным образом для проверки оставшихся условий выполнения регрессионного анализа. Одно из таких условий (№7) – нормальное распределение остатков с M=0. SPSS предлагает оценить нормальность распределения остатков графическим способом. На рис. 17 и 18 представлены гистограмма и квантильная диаграмма остатков, соответственно. Кроме того, на гистограмме указаны значения M = 1,39E-15, что равно $1{,}39/10^{15}$, стандартное отклонение и объем выборки. Мы видим, что распределение остатков очень близко к нормальному, а M к нулю.

Более точно нормальность распределения остатков можно проверить с помощью уже известных нам критериев Колмогорова-Смирнова и Шапиро-Уилка [3-5, 15], но для этого сначала нужно рассчитать эти остатки для каждого наблюдения, что делается SPSS автоматически непосредственно при проведении регрессионного анализа, если в меню «Save» выбрать unstandardized residuals (рис. 7). В результате в базе SPSS появляется новая переменная «REZ», нормальность распределения которой и необходимо проверить. При проверке соблюдения этого условия для нашего примера (Analyze, затем Descriptive statistics, затем Explore, где нужно перенести переменную «REZ» в окно «Dependent List», а в меню «Plots» отметить Histogram и Normality plots with test, вернуться назад нажатием на «continue» и запустить анализ нажатием на «OK») мы видим, что среднее арифметическое (M) равно 0,00 (рис. 19), медиана несколько больше, чем среднее, но модули коэффициентов асимметрии и эксцесса менее 1. Согласно тесту Kolmogorov-Smirnov с поправкой Lillefors [16] распределение не отличается от нормального, в то время как тест Shapiro-Wilk говорит об обратном (рис. 20). Критерий Шапиро-Уилка является наиболее мощным из тестов для проверки нормальности распределения и при значительном объеме выборки может быть чувствительным даже к минимальным отклонениям, которые не должны являться препятствием к применению множественного регрессионного анализа, поэтому рекомендуется также строить графики и визуально оценивать распределение остатков [3]. Посмотрев на гистограмму и квантильную диаграмму (рис. 21), мы видим, что распределение остатков незначительно отличается от нормального. Таким образом, это условие регрессионного анализа также можно считать выполненным.

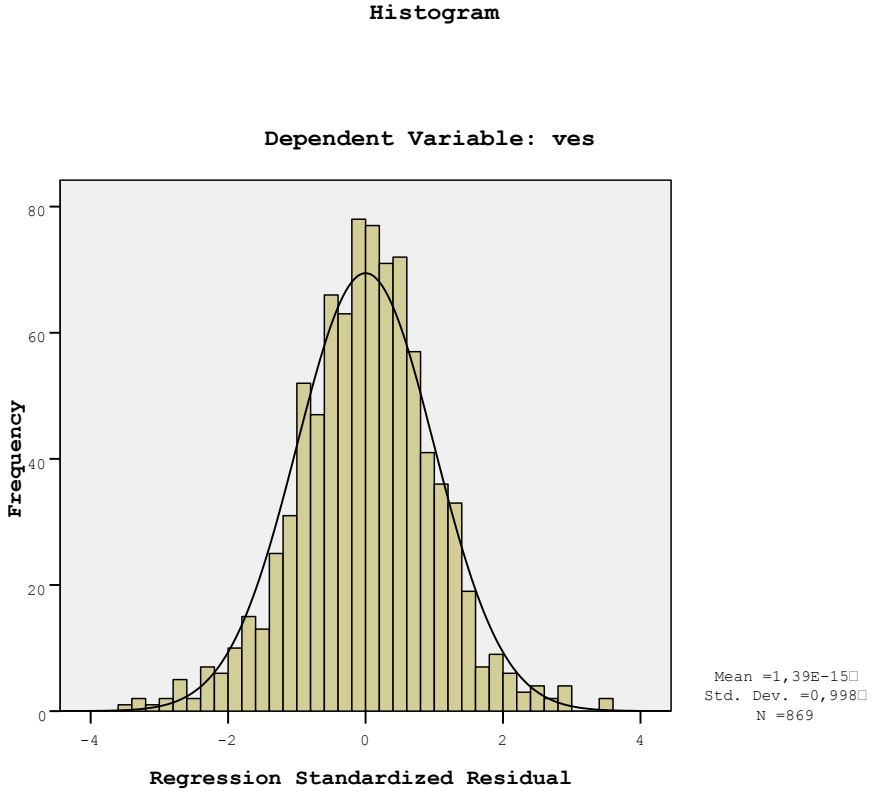

Рис. 17. Гистограмма, отражающая распределение стандартизованных остатков с кривой нормального распределения.

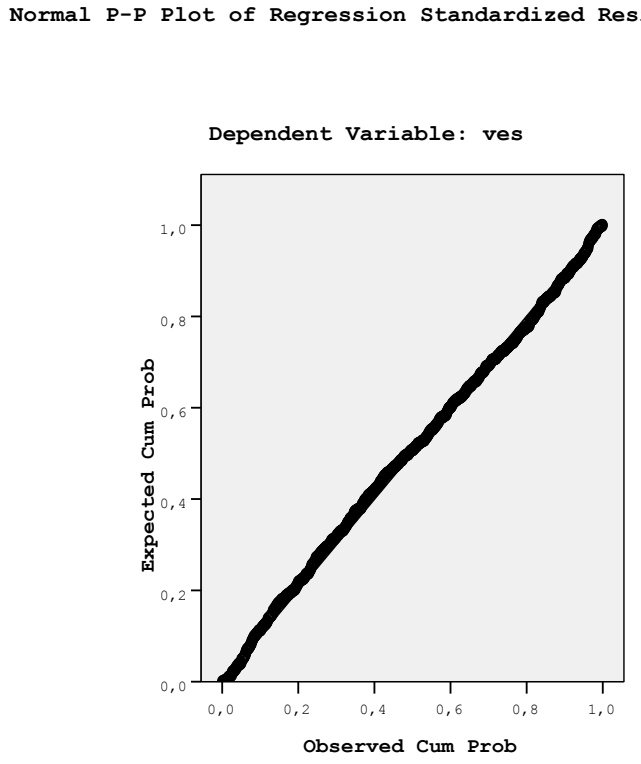

Рисунок 18. Квантильная диаграмма остатков.

Descriptives

			Statistic	Std. Error
Unstandardized Residual	Mean		,0000000	8,09406777
	95% Confidence Interval for Mean	Lower Bound	-15,8862330	
		Upper Bound	15,8862330	
	5% Trimmed Mean		1,9280075	
	Median		4,4097743	
	Variance		56931,608	
	Std. Deviation		238,60345303	
	Minimum		-818,48775	
	Maximum		843,95542	
	Range		1662,44318	
	Interquartile Range		292,99379	
	Skewness		-,112	,083
	Kurtosis		,774	,166

Рисунок 19. Результаты описательной статистики для переменной «REZ».

Tests of Normality

	Kolmogorov-Smirnov(a)			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Unstandardized Residual	,027	869	,155	,993	869	,000

a Lilliefors Significance Correction

Рисунок 20. Результаты тестов на нормальность распределения для переменной «REZ».

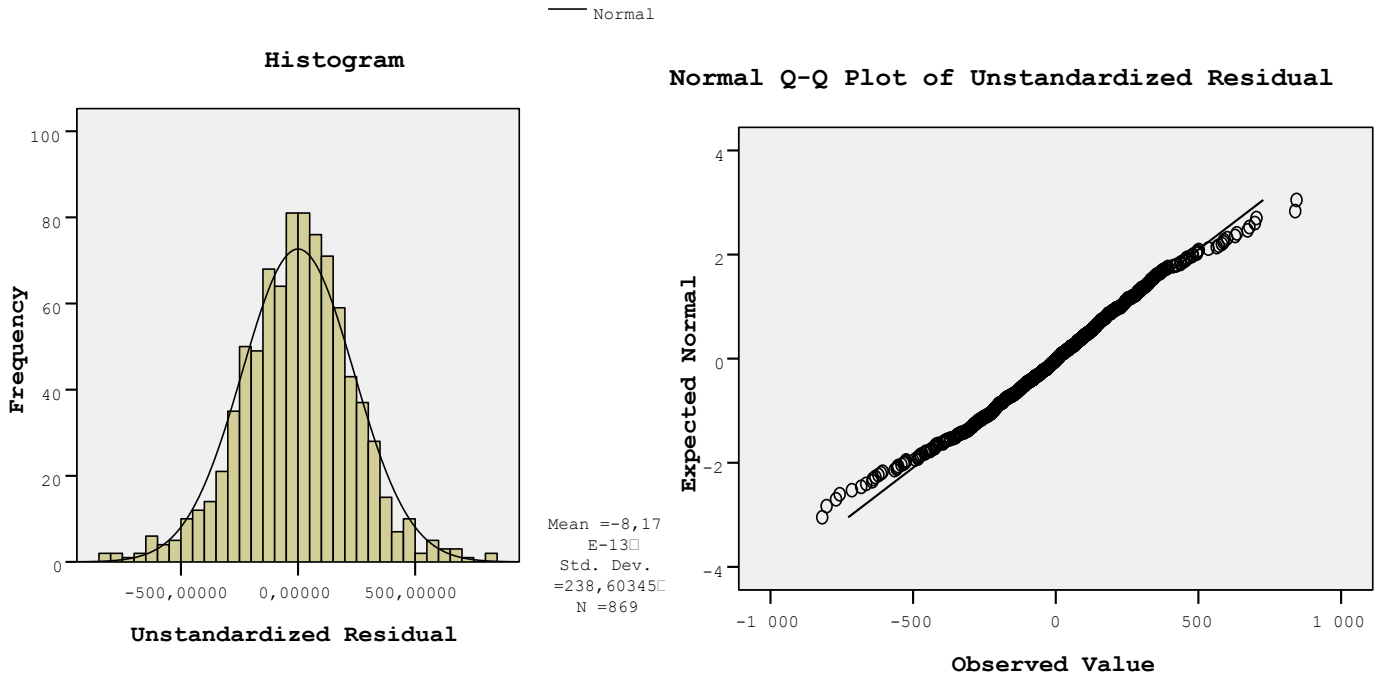

Рисунок 21. Гистограмма (слева) и квантильная диаграмма (справа) переменной «REZ».

Следующий график (рис. 22), или скаттерограмма, показывает, как распределены остатки в зависимости от предсказываемых значений зависимой переменной. Она необходима для проверки восьмого условия или наличия гомоскедастичности. В идеале график должен представлять собой как можно более дезорганизованный разброс точек, напоминая облако, но не должен иметь форму треугольника, двух облаков и т.д., что означало бы наличие гетероскедастичности [3-5]. В таком случае предсказательная способность регрессионной модели имела бы различную степень точности при различных уровнях зависимой переменной. Для нашего исследования это последнее условие также соблюдено.

Затем программой представляются три скаттерограммы, отражающие зависимость между переменной отклика и каждым из предикторов. Мы уже оценивали такие скаттерограммы для всех независимых переменных, строив их с помощью меню «Graphs», когда определяли линейность их взаимосвязи с массой тела, поэтому повторно их представлять не будем. Таким образом, можно избежать предварительного построения данных скаттерограмм «вручную» перед проведением регрессионного анализа, а проверить наличие линейной взаимосвязи переменной отклика с каждой из независимых переменных, непосредственно выполнив его.

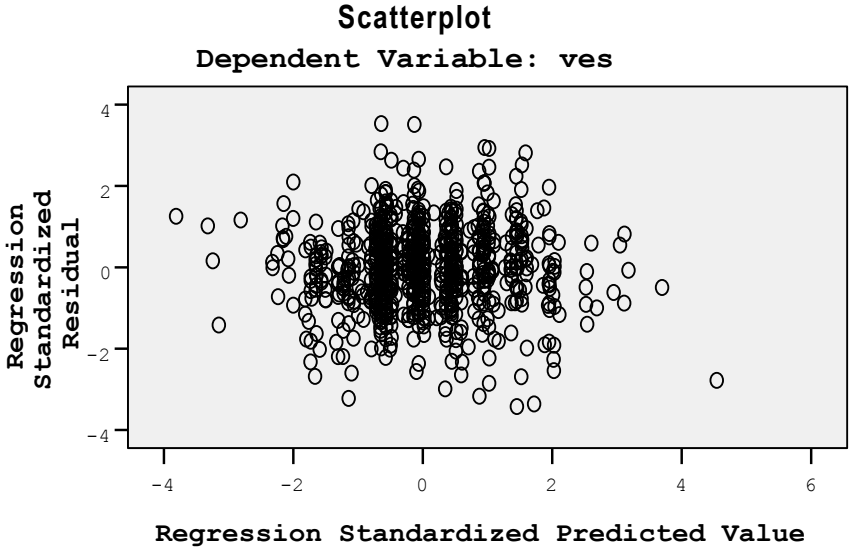

Рисунок 22. Разброс стандартизованных остатков в зависимости от стандартизованных предсказанных значений.

Еще один важный вопрос: что делать, если предиктор представлен в виде категориальной переменной? Можно ли включить, например, пол новорожденного в качестве еще одной независимой переменной в регрессионный анализ в нашем исследовании? Это возможно, причем, если предиктор имеет только две возможные категории (бинарная переменная), как пол в нашем примере, то нужно всего лишь закодировать их в виде 0 и 1 и включить в таком виде в анализ. Когда переменные имеют более, чем две категории, например уровень образования (таб. 1), процедура введения их в модель несколько усложняется. В таких случаях необходимо создавать дамми-переменные (dummy variables), что также является способом кодирования категориальной переменной через нули и единицы [17]. В результате мы получим несколько дамми-переменных, а именно на одну меньше, чем количество категорий. Для того чтобы создать такие переменные, необходимо выполнить следующие шаги:

1. Определить количество дамми-переменных для категориального признака, вычитая один из количества категорий;

2. Определить так называемую контрольную или референс-категорию, с которой будут сравниваться все остальные (желательно, чтобы это была наибольшая по объему группа людей). Эта группа будет закодирована 0 во всех дамми-переменных;

3. Создать дамми-переменную для каждой из категорий, кроме контрольной, обозначив наличие данной категории 1, а отсутствие 0 (рис. 23, 24)

4. Включить дамми-переменные в регрессионный анализ.

Таблица 1.

Создание двух «Dummy» переменных для признака «Образование» с 3 категориями.

Образование, категории	«Dummy» переменные	
	Ср. специальное	Высшее
Среднее	0	0
Ср. специальное	1	0
Высшее	0	1

Для создания каждой дамми-переменной в SPSS необходимо использовать меню «Transform», затем выбрать «Transform into different variables», в результате чего откроется одноименное диалоговое окно (рис. 24). В левом поле необходимо выбрать необходимую категориальную переменную и перенести ее в правое поле, нажав на стрелку. Введя название соответствующей дамми-переменной под «Name», например «sred_spec» для категории «средне-специальное», и краткое описание этой новой переменной под «Label», например «дамми-переменная для категории средне-специальное образование», нажимаете на кнопку «Change». Затем, нажав на кнопку «Old and New Variables», открываете соответствующее окно (рис. 25). В левой половине окна (Old Value) необходимо последовательно вводить значения категорий исходной категориальной переменной, т.е. «образование», а в правой половине (New Value) – новое обозначение соответствующей категории в соответствии с принципом, описанным выше. Для дамми-переменной «sred_spec» категория «средне-специальное» должна принять значение 1, а остальные, т.е. «среднее» и «высшее» – нулями. После обозначения каждой категории нужно нажимать на кнопку «Add» для добавления ее в поле справа «Old→New». После того, как все категории для соответствующей дамми-переменной закодированы через нули и единицы, нажимаете на «Continue», а затем на «ОК». В базе SPSS появляется соответствующая дамми-переменная, для которой в разделе «Values» в режиме «Variable View» необходимо подписать, что было закодировано 1 (средне-специальное образование), а что 0 (среднее и высшее образование). Такую процедуру нужно выполнить для каждой дамми-переменной, т.е. в случае с образованием еще один раз (для высшего образования). Для среднего, как вы помните, дамми-переменная не создается, т.к. оно является референс-категорией.

В регрессионную модель в качестве предиктора вводится уже не переменная «образование» с тремя категориями, а две дамми-переменные: «средне-специальное образование» и «высшее образование». В результате мы получим регрессионные коэффициенты для каждой из дамми-переменных с достигнутым уровнем значимости, исходя из которых мы сможем определить, оказывает ли влияние на зависимую переменную соответствующая категория интересующей нас переменной по сравнению с референс-категорией или нет (уровень статистической значимости, p), и в каком направлении (знак регрессионного коэффициента). Также мы можем сравнить значимость влияния на переменную отклика каждой из категорий, кроме референс-категории, используя значения стандартизованных коэффициентов регрессии (чем модуль больше, тем более сильное влияние). Значение коэффициента само по себе интерпретировать достаточно сложно. Оно необходимо для представления регрессионной модели в виде уравнения, особенно если целью регрессионного анализа является прогнозирование.

Попробуем включить в нашу модель пол ребенка. Это единственная номинальная (а более точно бинарная) переменная, информация о которой была собрана в ходе Северодвинского исследования, и которая вполне может влиять на вес новорожденного. Также попробуем исключить переменную «vozrast», которая по результатам уже проведенного множественного регрессионного анализа не оказывала статистически значимого влияния на переменную отклика. Результаты представлены в таблице 2. Мы можем видеть, что мужской пол, который был закодирован единицей, не оказывает статистически значимого влияния на вес новорожденного, т.к. 95% доверительный интервал для коэффициента регрессии включает 0. Кроме того, добавление пола ребенка в качестве предиктора к длине и гестационному сроку не приводит к улучшению предсказательной способности модели (ΔR^2 <,001; p= ,469).

Таким образом, мы знаем, для чего и как проводить множественный линейный

регрессионный анализ с использованием SPSS и как правильно интерпретировать полученную в результате анализа информацию. Но очень важно перед тем, как докладывать результаты исследования, проверить, соответствует ли модель имеющимся выборочным данным или она подвержена влиянию ряда атипичных наблюдений. Это так называемая диагностика модели, которая включает в себя оценку остатков, «выскакивающих случаев» и случаев, оказывающих сильное влияние на модель. Диагностика модели производиться для оценки качества модели, а не для нахождения тех или иных наблюдений, исключение которых превратит недостоверные параметры в достоверные [18].

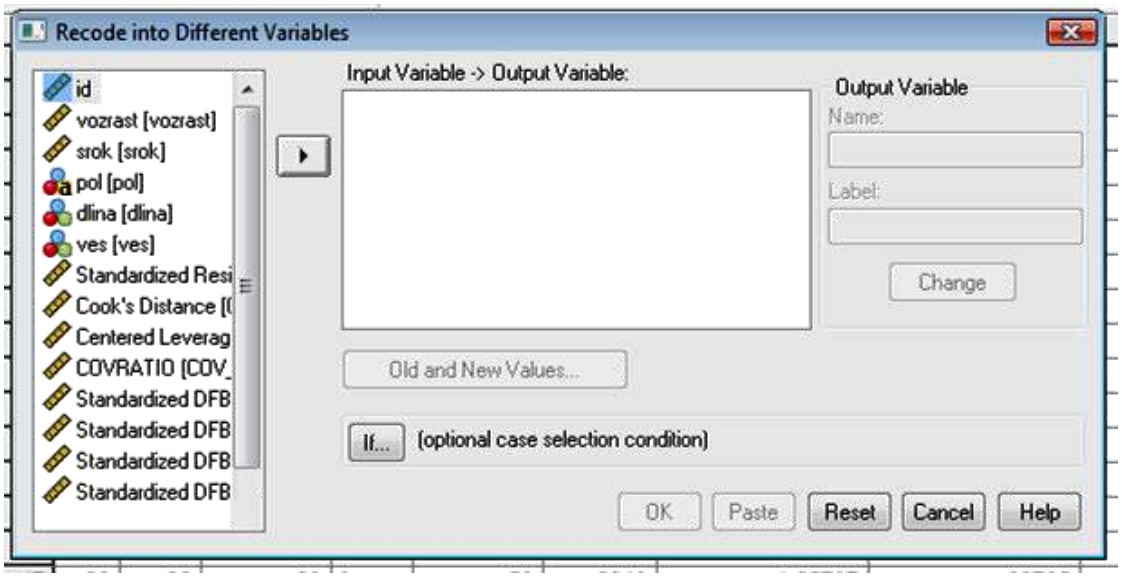

Рисунок 23. Диалоговое окно «Recode into Different Variables».

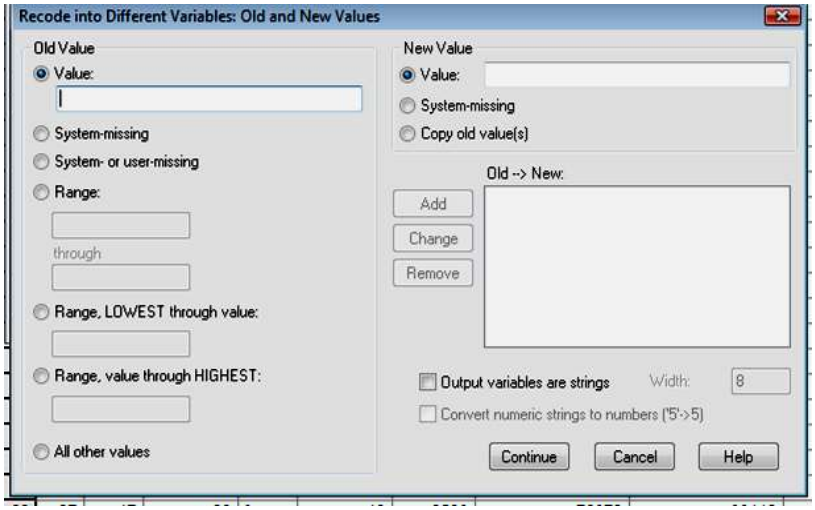

Рисунок 24. Диалоговое окно «Recode into Different Variables: Old and New Variables».

Второй важный вопрос, на который необходимо ответить: можно ли генерализировать (переносить) результаты регрессионного анализа, полученного на выборочных данных, на другие выборки и популяцию в целом? Для этого необходимо проверить соблюдение всех условий, необходимых для проведения множественной линейной регрессии (см. выше), а также проверить способность модели к генерализации путем ее кросс-валидации (cross-validation of the model). Первый способ кросс-валидации модели – сравнить скорректированный R^2 (adjusted R square) с грубым (R square). Как мы уже упоминали, скорректированный R^2 обычно меньше грубого, что указывает на потерю предсказательной способности модели, когда

она используется для всей популяции, а не для выборки, на основании которой была построена. Следовательно, чем больше разница между скорректированным и грубым значением, тем меньшей способностью к генерализации обладает модель. Второй возможный способ кросс-валидации – разбить случайным образом всю выборку на две части и сравнить результаты регрессионного анализа, выполненного для каждой из двух частей.

Важно помнить о том, что для построения надежной множественной линейной регрессионной модели важным является объем выборки. Какое же количество наблюдений будет достаточным? Существует множество правил, но одно из наиболее принятых – не менее 15-20 случаев на каждый предиктор [3-4, 10]. Например, в нашем случае было 3 предиктора, следовательно объем выборки должен быть не менее 45-60 человек. Большое значение для определения необходимого минимального объема выборки имеет также размер эффекта, который мы пытаемся выявить, и размер статистической мощности, при этом иногда будет достаточно порядка 10 наблюдений на одну независимую переменную [3, 10]. Но в любом случае, чем больше будет выборка, тем более надежной будет модель, и тем большей способностью к генерализации она будет обладать.

При представлении результатов множественного линейного регрессионного анализа необходимо указывать нескорректированные (полученные при проведении простой линейной регрессии) и скорректированные (полученные в результате множественного регрессионного анализа) регрессионные коэффициенты с 95% доверительными интервалами (что предпочтительнее) или с уровнем их статистической значимости (p). Также рекомендуется приводить коэффициент детерминации (R^2) и константу (b_0), особенно если модель используется с предсказательной целью. В случаях, когда при анализе использовался пошаговый или иерархический метод ввода независимых переменных, желательно представлять эти данные для каждой из моделей, а также приводить значение R^2 change с уровнем статистической значимости для изменения. Нагляднее, когда результаты представлены в виде таблицы (табл. 2, 3).

На что следует обратить внимание при оформлении таблиц? Значения всех показателей округлены до одинакового количества знаков после запятой, до 2 в нашем примере. Стандартизованные регрессионные коэффициенты и коэффициенты детерминации не содержат 0 до запятой, т.к. их значения не могут превышать 1. Общая информация, такая как R^2, ΔR^2 и т.д., представлена под таблицами в примечаниях. Кроме того, хотелось бы отметить, что не нужно дублировать информацию. Другими словами, если результаты регрессионного анализа представлены в виде таблиц, то в тексте приводятся лишь комментарии.

Таблица 2.

Предикторы массы тела новорожденного в г. Северодвинск по результатам множественного линейного регрессионного анализа (n=869).

Признаки	b	95% ДИ для b	β (стаднарт. b)
Блок 1			
Константа (b_0)	-7129,15	-7700,39; -6557,92	
Длина ребенка, см	183,75	174,94; 192,56	,81
Срок гестации, нед.	28,06	14,76; 41,46	,08
Блок 2			
Константа (b_0)	-7123,02	-7694,66; -6551,39	
Длина ребенка, см	183,17	174,23; 192,12	,80
Срок гестации, нед.	28,49	15,14;41,85	,08
Пол ребенка, мужской	11,95	-20,39; 44,29	,01

Примечания. R^2 = ,70 для блока 1 (p< ,001); ΔR^2 <,001 для блока 2 (p= ,469).

Таблица 3.

Представление результатов множественного линейного регрессионного анализа.

Признаки	Однофакторный анализ[a]		Многофакторный анализ[b]	
	b (95%ДИ)	p	b (95%ДИ)	p
Длина ребенка, см	189,83 (181,43; 198,23)	<0,001	183,85 (175,03; 192,67)	<0,001
Срок гестации, нед.	118,95 (97,43; 140,48)	<0,001	28,14 (14,83; 41,46)	<0,001
Возраст матери, года	6,15 (-1,78; 14,09)	0,128	-0,91 (-5,28; 3,46)	0,684

Примечания.

[a] – коэффициенты регрессии с достигнутым уровнем статистической значимости по результатам простого линейного регрессионного анализа.

[b] – коэффициенты регрессии с достигнутым уровнем статистической значимости по результатам множественного линейного регрессионного анализа;

b0 = -7116,96; R^2 скор.= ,70, p<0,001.

Таким образом, в данной статье мы постарались представить основные этапы проведения и интерпретации метода множественной линейной регрессии и дать ответы на наиболее часто возникающие вопросы при использовании данного вида анализа.

Литература:

1. *Гржибовский А.М.* Однофакторный линейный регрессионный анализ // Экология человека. 2008. №10. С. 55-64.
2. *Гржибовский А.М., Иванов С.В.* Однофакторный линейный регрессионный анализ с использованием программного обеспечения Statistica и SPSS // Наука и Здравоохранение 2017. №2. С. 5-33.
3. *Юнкеров В.И., Григорьев С.Г.* Математико-статистическая обработка данных медицинских исследований. СПб.: ВМедА, 2002. 266 с.
4. *Field A.* Discovering statistics using SPSS (2nd ed.). London: Sage Publications Ltd., 2005. 781 p.
5. *Foster J.* Understanding and using advanced statistics. Foster J., Barkus M., Yavorsky C. London: SAGE Publications Ltd., 2006. 178 p.
6. *Grjibovski A. M., Bygren L.O., Svartbo P.M.* Social variations in fetal growth in Northwest Russia: an analysis of medical records // Annals of Epidemiology. 2003. N 9. pp. 599-605.
7. *Grjibovski A.M., Bygren L.O., Yngve A., Sjostrom M.* Social variations in infant growth performance in Severodvinsk, Northwest Russia: community-based cohort study // Croat Med J. 2004. V. 45. N 6. pp. 757-63.
8. *Grjibovski A., Bygren L.O., Svartbo B. Magnus P.* Housing conditions, perceived stress, smoking, and alcohol: determinants of fetal growth in Northwest Russia // Acta Obstet Gynecol Scand. 2004. V. 83. N 12. pp. 1159-66.
9. *Grjibovski A., Bygren L.O., Svartbo B.* Socio-demographic determinants of poor infant outcome in north-west Russia // Paediatr Perinat Epidemiol. 2002. V. 16. N 3. pp. 255-62.
10. *Little R. J. A.* A test of missing completely at random for multivariate data with missing values. // Journal of the American Statistical Association.1998. N 83. P. 1198-1202.
11. *Brooks G.P., Barcikowski R.S.* The PEAR method for sample sizes in multiple linear regression // Multiple Linear Regression Viewpoints. 2012. V. 38. N 2. pp. 1-16.
12. *Cook R. D.*, *Weisberg S*. Residuals and influence in regression. New York – London: Chapman and Hall, 1982. 229 p.
13. *Hoaglin D.C., Welsch R.E.* The Hat Matrix in Regression and ANOVA // the American statistician. 1978. V. 32. N 1. P. 17–22.
14. *Stevens J.P*. Applied Multivariate Statistics for the Social Sciences using SAS & SPSS (4th ed.). New York: Psychology Press, 2002. 708 p.
15. *Shapiro S.S., Wilk M.B.* An analysis of variance test for normality // Biometrika, 1965. V. 52. N 3. P. 591-611.
16. *Lilliefors H.* On the Kolmogorov–Smirnov test for normality with mean and variance unknown // Journal of the American Statistical Association. 1967. N 62. pp. 399–402.
17. *Suits D.B.* Use of Dummy Variables in Regression Equations // Journal of the American Statistical Association. 1957. V. 52. N 280. pp. 548–551.

18. *Belsey D.A., Kuh E., Welsch R.E.* Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: John Wiley and Sons. 1980. 300 p.

References:

1. Grjibovski A.M. Odnofactornyj lineinyj regressionnyj analiz. [Simple linear regression analysis*]. Ekologiya cheloveka* [Human ecology] 2008, 10, pp. 55-64. [in Russian].

2. Grjibovski A.M., Ivanov S.V., Gorbatova M.A. Odnofaktornyi lineinyi regressionnyi analiz s ispol'zovaniem programmnogo obespecheniya Statistica i SPSS [Univariate regression analysis using Statistica and SPSS software]. *Nauka i Zdravookhranenie* [Science & Healthcare]. 2017. 2, pp. 5-33. [in Russian].

3. Junkerov V.I., Grigoriev S.G. *Matematiko-statisticheskaya obrabotka dannykh medtscinskikh issledovanii* [Mathematical ans statistical analysis of the medical research data]. SPb: VMedA, 2002. 266 p. [in Russian].

4. Field A. *Discovering statistics using SPSS (2nd ed.).* London: Sage Publications Ltd., 2005. 781 p.

5. Foster J. *Understanding and using advanced statistics.* Foster J., Barkus M., Yavorsky C. London: SAGE Publications Ltd., 2006. 178 p.

6. Grjibovski A.M., Bygren L.O., Svartbo P.M. Social variations in fetal growth in Northwest Russia: an analysis of medical records. *Annals of Epidemiology*. 2003, 9, p.p. 599-605.

7. Grjibovski A.M., Bygren L.O., Yngve A., Sjostrom M. Social variations in infant growth performance in Severodvinsk, Northwest Russia: community-based cohort study. *Croat Med J.* 2004, 45(6), p.p. 757-63.

8. Grjibovski A., Bygren L.O., Svartbo B. Magnus P. Housing conditions, perceived stress, smoking, and alcohol: determinants of fetal growth in Northwest Russia. *Acta Obstet Gynecol Scand.* 2004, 83(12), p.p. 1159-66.

9. Grjibovski A., Bygren L.O., Svartbo B. Socio-demographic determinants of poor infant outcome in north-west Russia*. Paediatr Perinat Epidemiol*. 2002, 16(3), p.p. 255-62.

10. Little R. J. A. A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association.* 1998, 83, p.p. 1198-1202.

11. Brooks G.P., Barcikowski R.S. The PEAR method for sample sizes in multiple linear regression. *Multiple Linear Regression Viewpoints.* 2012, 38(2), p.p. 1-16.

12. Cook R.D., Weisberg S. *Residuals and influence in regression.* New York – London: Chapman and Hall, 1982. 229 p.

13. Hoaglin D.C., Welsch R.E. The Hat Matrix in Regression and ANOVA. The American statistician. 1978, 32(1), p.p. 17–22.

14. Stevens J.P. Applied *Multivariate Statistics for the Social Sciences using SAS & SPSS (4th ed.)*. New York: Psychology Press, 2002. 708 p.

15. Shapiro S.S., Wilk M.B. An analysis of variance test for normality. *Biometrika.* 1965, 52(3), p.p. 591-611.

16. Lilliefors H. On the Kolmogorov–Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association.* 1967, 62, p.p. 399–402.

17. Suits D.B. Use of Dummy Variables in Regression Equations. *Journal of the American Statistical Association.* 1957, 52(280), p.p. 548–551.

18. Belsey D.A., Kuh, E., Welsch, R.E. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley and Sons. 1980. 300 p.

Контактная информация:

Гржибовский 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. Ясяви, г, Туркестан, Казахстан; Почетный профессор ГМУ г. Семей, Казахстан.

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

Email: Andrej.Grjibovski@gmail.com

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