Получена: 15 июля 2017 / Принята: 18 августа 2017 / Опубликована online: 30 августа 2017

УДК 614.2 + 303.4

ПРИМЕНЕНИЕ МНОЖЕСТВЕННОГО ЛОГИСТИЧЕСКОГО 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 логистического регрессионного анализа для оценки связи между одной зависимой дихотомической переменной и нескольких независимых переменных с использованием пакета прикладных статистических программ SPSS. Также подробно описаны принципы интерпретации полученной информации на практическом примере.

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

Abstract

MULTIVARIABLE LOGISTIC REGRESSION USING SPSS SOFTWARE IN HEALTH RESEARCH

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 logistic regression analysis for predicting one dichotomous outcome based on several independent variables using the SPSS software. The article describes basic principles and peculiarities of interpretation of the results using practical examples. We also describe advantages and disadvantages of this type of analysis

Key words: *multivariable logistic regression analysis, coefficient of determination, least squares distance method, confidence intervals, SPSS.*

Түйіндеме

SPSS СТАТИСТИКАЛЫҚ БАҒДАРЛАМАЛАР ПАКЕТІН ПАЙДАЛАНУМЕН ДЕНСАУЛЫҚ САҚТАУДАҒЫ КӨПШІЛІК ЛОГИСТИКАЛЫҚ 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 бір тәуелді дихотомиялық ауыспалы және бірнеше тәуелді емес ауыспалыларды пайдаланумен арасындағы байланысты бағалау үшін көптеген логистикалық регрессивтік талдауды өткізудің теориялық негіздері берілген. Сол сияқты толық осы әдісті қолдану кезінде шыққан негізгі мәселелер анықталды және оларды шешудің нұсқалары ұсынылған.

Негізгі сөздер: *көпшілік логистикалық 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 логистического регрессионного анализа в здравоохранении с использованием пакета статистических программ SPSS / / Наука и Здравоохранение. 2017. №4. С. 5-26.

Sharashova E.E., Kholmatova K.K., Gorbatova M.A., Grjibovski A.M. Application of the multivariable logistic regression analysis in healthcare using SPSS software. *Nauka i Zdravookhranenie* [Science & Healthcare]. 2017, 4, pp. 5-26.

Шарашова Е.Е., Холматова К.К., Горбатова М.А., Гржибовский А.М. SPSS статистикалық бағдарламалар пакетін пайдаланумен денсаулық сақтаудағы көпшілік логистикалық регрессивтік талдауды қолдану / / Ғылым және Денсаулық сақтау. 2017. №4. Б. 5-26.

Из предыдущих статей и выпусков данного журнала [1, 2, 4] мы знаем, что использование линейного регрессионного анализа позволяет нам прогнозировать значение зависимой переменной по известным значениям одной или нескольких переменных-предикторов. Но одним из ключевых условий, необходимых для выполнения линейного регрессионного анализа, является количественный, а точнее интервальный характер зависимой переменной. В тоже время, существует множество ситуаций, когда переменная отклика, значение которой мы бы хотели предсказать на основании тех или иных предикторов, является бинарной (дихотомической). Например, как ответить на вопрос, какие из имеющихся переменных влияют на вероятность пациента умереть (зависимая переменная бинарная – умер/не умер), или влияет ли назначение какого-либо препарата на вероятность пациента поправиться (зависимая переменная – поправился/не поправился), или какова вероятность того, что опухоль, выявленная у пациента, злокачественная (зависимая переменная – злокачественная / доброкачественная) и т.д? В таких ситуациях логистический регрессионный анализ является анализом выбора. Множественный логистический регрессионный анализ дает

возможность анализировать взаимосвязь между бинарной переменной отклика (зависимой переменной) и любыми, количественными или качественными, переменными-предикторами (независимыми переменными); позволяет прогнозировать, к какой из двух групп принадлежит изучаемый случай в зависимости от известных значений переменных-предикторов. Таким образом, логистическая регрессия – это та же множественная регрессия с той лишь разницей, что зависимая переменная категориальная, а независимые переменные могут быть любыми.

Основные принципы логистической регрессии [3, 5, 7, 10, 11, 20]. В простой линейной регрессии, для того, чтобы предсказать значение зависимой переменной мы использовали линейную модель, или 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, или значение y, в котором прямая линия пересекает ось y,

b_1 – регрессионный коэффициент, или угол наклона прямой линии,

ε_i – случайная ошибка.

На основании значений Y_i и X_i, полученных на выборке из интересующей нас популяции, можно, используя метод наименьших квадратов, т.е. минимизируя квадраты разниц между фактическими и предсказываемыми значениями зависимой переменной, рассчитать значения неизвестных параметров (b_0 и b_1). В результате мы получим простую линейную регрессионную модель, которую можно использовать для предсказания значения Y для любого члена исходной популяции по известному значению X_i. Все это Вам уже знакомо из предыдущих выпусков.

Аналогично сторонится и множественная линейная регрессионная модель. В этом случае уравнение включает не один, а несколько предикторов, каждый из которых имеет свой регрессионный коэффициент:

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

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).

В логистической же регрессии на основании значений одной или нескольких переменных-предикторов мы предсказываем не значение зависимой переменной Y, как это было в линейной регрессии, а вероятность принадлежности индивидуума к одной из двух категорий переменной Y. Уравнение логистической регрессии во многом схоже с только что представленным:

$$P(Y) = 1 / 1 + e^{-(b0 + b1 \cdot X1i + b2 \cdot X2i + \cdots + bn \cdot Xni + \varepsilon i)},$$

где P(Y) – вероятность принадлежность к одной из двух категорий зависимой переменной,

e – основание натурального логарифма (e $\approx$ 2,72),

$b_0 + b_1 \cdot x_{1i} + b_2 \cdot x_{2i} + \cdots + b_n \cdot X_{ni} + \varepsilon_i$ – правая часть уравнения множественной линейной регрессии, где коэффициенты несут ту же самую информацию.

Несмотря на видимое сходство моделей, лежащих в основе линейной и логистической регрессий, мы не можем использовать уравнение линейной регрессии в ситуациях, когда зависимая переменная дихотомическая. Почему? Одним из условий, необходимых для выполнения линейного регрессионного анализа, является наличие линейной взаимосвязи между зависимой и независимой переменными. Если зависимая переменная дихотомическая, это условие изначально не может быть соблюдено. Именно это и лежит в основе различий между линейным и логистическим уравнениями: последнее является логистической трансформацией первого. Другими словами, уравнение логистической регрессии представляет собой уравнение линейной регрессии на логарифмической шкале. Логарифмическая трансформация позволяет выразить нелинейную взаимосвязь в форме линейной.

Как мы уже отмечали, для выполнения множественного линейного регрессионного анализа требуется соблюдение целого ряда условий [4]. При проведении множественного

логистического регрессионного анализа количество условий меньше, но они все-таки существуют:

1. зависимая переменная должна быть дихотомической;
2. желаемый исход в зависимой переменной должен быть закодирован единицей, так как в логистической регрессии вероятность возникновения события принимается за единицу (P(Y=1));
3. независимость наблюдений;
4. отсутствие мультиколлинеарности, т.е. ситуаций, когда независимые переменные сильно коррелируют между собой (r >0,9);
5. линейная зависимость между каждой независимой переменной и логарифмом отношения шансов (log odds);
6. независимость остатков.

Выполнение условий нормальности распределения остатков, гомоскедастичности, основанных на методе наименьших квадратов, при проведении линейного регрессионного анализа не требуется.

Однако, при линейном регрессионном анализе необходим больший объем выборки. Требуется большее (по сравнению с линейной регрессией) количество наблюдений на одну независимую переменную (от 30 наблюдений), так как показатель log-likelihood менее мощный по сравнению с расчетом наименьших квадратов.

Значение зависимой переменной в уравнении логистической регрессии, P (Y), может принимать любое значение от 0 до 1, при этом значения близкие к 0 – означают, что вероятность индивидуума принадлежать к одной из категорий зависимой переменной (например, вероятность умереть, если зависимая переменная умер/не умер) крайне мала, а близкие к 1 – что эта вероятность крайне велика. Как и в линейной регрессии, каждый предиктор в логистическом регрессионном уравнении имеет свой коэффициент, а найти эти коэффициенты (b_1, b_2, … b_n), также как и константу (b_0), и является основной целью проведения анализа. SPSS подбирает значения указанных коэффициентов, и оставляет в результате те, при которых получившаяся модель наиболее точно отражает наши фактические данные. В итоге мы получаем ту модель, которая при включении в нее всех рассчитанных параметров дает значения P(Y) наиболее близкие к эмпирическим (Y).

Как же мы можем оценить качество нашей модели. Для этого необходимо посмотреть, насколько точно она соответствует фактическим данным. Используя полученное уравнение логистической регрессии, мы можем рассчитать для каждого индивидуума в нашей выборке вероятность возникновения события, или, другими словами, вероятность принадлежать к одной из двух категорий зависимой переменной (например, вероятность умереть). И эта вероятность, P(Y), может принимать любое значение от 0 до 1. Фактические же данные, на основании которых SPSS и строила модель, содержат точную информацию для каждого индивидуума о том, произошло событие или нет (например, умер или нет), т.е. Y равное либо 0, либо 1. Для того чтобы оценить модель, а именно ее предсказательную способность, необходимо сравнить предсказываемые значения переменной отклика с фактическими. В линейной регрессии для этих целей мы использовали коэффициент детеминации, R^2, который равен квадрату коэффициента корреляции между предсказанными и фактическими значениями переменной отклика. В логистической регрессии мы используем показатель log-likelihood:

$$\text{log-likelihood} = \sum_{i=1}^{N} \{Y_i \ln(P(Y_i)) + (1-Y_i) \ln[1-P(Y_i)]\}.$$

Показатель log-likelihood является аналогом суммы квадратов остатков в линейной регрессии (SS_R). Он показывает, сколько необъясненной информации осталось после использования модели для фактических данных. Следовательно, чем больше значение показателя, тем хуже модель предсказывает имеющиеся данные. Но где же граница между плохой моделью, и той, которую мы можем использовать в дальнейшем?

В линейной регрессии мы сравнивали построенную модель с простейшей, в качестве

которой использовали среднее значение переменной отклика (Y). В логистической регрессии в качестве базовой, или простейшей модели используется то значение зависимой переменной, Y, которое чаще встречается в выборке. Например, если в выборке из 100 человек умерло 72, а 28 остались живы, то базовая модель предсказывала бы для каждого индивидуума из этой популяции смертельный исход. Другими словами, если бы мы не имели никаких других данных (наши предикторы), то для того, чтобы предсказать исход для какого-либо индивидуума (например, умрет или нет), мы бы использовали тот вариант, который произошел у большинства. Таким образом, мы можем рассчитать значения log-likelihood для оценки каждой из моделей, логистической и базовой, сравнить их и узнать, повышает ли наша модель (т.е. добавление тех или иных предикторов) предсказательную способность базовой модели (содержит только константу: 0 или 1) или нет, а также значимо ли это улучшение или нет. Для этого рассчитывается показатель хи-квадрат (χ^2):

$$\chi^2 = 2\ (LL(\text{нов. модель}) - LL(\text{Базовая модель})),\ df = k_{Нов.} - k_{Баз.}$$

Умножение правой части уравнения на 2 необходимо, чтобы привести значение разности к распределению χ^2, а это в свою очередь позволяет рассчитать статистическую значимость. Распределение хи квадрат имеет количество степеней свободы равное разности между количеством параметров в новой модели ($k_{Нов.}$), которое равно количество предикторов плюс 1 (константа), и количеством параметров в базовой модели ($k_{Баз}$), которое всегда равно 1, т.к. эта модель содержит только один параметр – константу. Если значение χ^2 для модели выходит за пределы критического значения, которое определяется соответствующим количеством степеней свободы, то можно говорить, что при определенном уровне значимости модель предсказывает исход статистически значимо лучше, чем базовая модель. Это значит, что хотя бы один из предикторов, включенных в модель статистически значимо влияет на вероятность исхода.

Коэффициент детерминации (R^2) в линейной регрессии позволял судить какой процент вариабельности зависимой переменной объясняется вариабельностью независимых. Значение показателя log-likelihood так нельзя интерпретировать. Он лишь позволяет судить о статистической значимости модели. Помимо log-likelihood (-2LL) SPSS рассчитывает и два аналога R^2 с использованием формулы Cox & Snell (1989) [9]:

$$R^2_{CS} = 1 - e^{[-2/n\ (LL(Нов.) - LL(Баз.))]}$$

и формулы Nagelkerke (1991) [18]:

$$R^2_N = R^2_{CS} / [1 - e^{(-2\ (LL(Баз.))}].$$

R^2, рассчитанный по формуле Cox & Snell, не может достичь своего теоретического максимума, т.е. 1, или 100%, поэтому предпочтительнее использовать второй вариант коэффициента (R^2_N). Кроме того, существует еще один более простой вариант расчета R^2 для логистической регрессионной модели (Hosmer & Lemeshow, 1989) [10]:

$$R^2_L = \chi^2 \text{ итоговой модели} / -2 \text{ Log likelihood базовой модели}$$

Несмотря на то, что существует несколько вариантов расчета коэффициента детерминации для логистической регрессионной модели, его значение интерпретируется одинаково, и подобно тому, как это делается в линейной регрессии.

Помимо оценки качества модели в целом SPSS, позволяет оценить вклад в предсказательную способность каждого предиктора в отдельности и независимо друг от друга. В линейной регрессии для этой цели мы использовали регрессионный коэффициент (b) и критерий Стьюдента для оценки его статистической значимости. Аналогичная процедура проводится и при выполнении логистического регрессионного анализа. В логистической регрессии нулевая гипотеза о том, что предиктор никак не связан с зависимой переменной, т.е. регрессионный коэффициент не отличается от 0 (H_0: b=0),

проверяется с помощью критерия Wald. Если регрессионный коэффициент статистически значимо отличается от 0, т.е. при определенном уровне значимости нулевая гипотеза отвергается ($b \neq 0$), то предиктор вносит статистически значимый вклад в предсказательную способность модели.

Регрессионный коэффициент в логистической регрессии необходим для оценки статистической значимости предиктора, но сложен для интерпретации сам по себе. Так, на основании этого значения, мы можем сказать, что тот или иной предиктор статистически значимо взаимосвязан, или не взаимосвязан с переменной отклика. Но если взаимосвязь статистически значима, то какова она? Значительно больше информации о степени и характере взаимосвязи предиктора с зависимой переменной несет значение коэффициента Exp(B). Этот коэффициент показывает во сколько раз изменяются шансы возникновения события (например, шансы умереть, если зависимая переменная умер/не умер), при изменении значения независимой переменной на единицу. Например, мы хотим посмотреть, влияет ли и как назначение лечения (независимая переменная) на вероятность пациента умереть (зависимая переменная). Шансы, что событие произойдет, определяется как отношение вероятности возникновения события (вероятность умереть) к вероятности того, что событие не произойдет (вероятность не умереть):

шансы (odds) = P (событие Y произошло) / P (событие Y не произошло),
где P (событие Y произошло) = $1 / [1+e^{-(b0+b1x1)}]$, а
P (событие Y не произошло) = 1 – P (событие Y произошло).

Для того, чтобы рассчитать во сколько раз изменятся шансы умереть при изменении предиктора на единицу (т.е. в нашем примере, при наличии лечения (1) по сравнению с отсутствием лечения (0)), необходимо сначала рассчитать шансы умереть для тех, у кого лечение проводилось, затем для тех, кто лечения не получал. Разделив первый показатель на второй, мы получим нужное значение – отношение шансов (Odds Ratio). Значение Exp(B), то есть отношения шансов, больше единицы говорит о том, что связь между предиктором и зависимой переменной положительная, т.е. увеличение значения предиктора увеличивает вероятность возникновения события. Значение Exp(B) менее единицы означает, что увеличение значения предиктора уменьшает шансы возникновения события [10].

Вы помните, что при проведении множественного линейного регрессионного анализа в SPSS, мы могли использовать несколько методов ввода независимых переменных в модель [4]. При проведении логистического регрессионного анализа доступны несколько из них: метод форсированного ввода, Enter (все переменные вводятся в модель одновременно, одним или несколькими блоками), и пошаговые методы (последовательного ввода, forward, и последовательного исключения, backward). Метод форсированного ввода используется SPSS по умолчанию и, по мнению многих исследователей, является единственно правильным для проверки теории, т.к. пошаговые методы подвержены влиянию случайных вариаций и поэтому редко приводят к получению воспроизводимых моделей [10]. Однако в ситуациях, когда подобных исследований не проводилось, и нет данных, на которые можно бы было опереться и построить гипотезу, а также, когда основная цель построить модель с максимальной предсказательной способностью, а не изучить взаимосвязи между переменными, применение пошаговых методов может быть оправдано [17].

При пошаговых способах введения переменных, которые подробно были описаны ранее [4], исследователь самостоятельно только выбирает ряд интересующих его предикторов, а программа, основываясь исключительно на математических критериях, определяет, в каком порядке они будут вводиться в модель, и какие из них останутся в модели в итоге. На каждом этапе, как метода последовательного ввода, так и метода последовательного исключения производится оценка очередного предиктора, на основании которой предиктор либо остается в модели, либо нет. SPSS предлагает по три варианта каждого из пошаговых методов: LR, Conditional

и Wald, которые и отличаются друг от друга как раз способом оценки очередного предиктора, а точнее его вклада в предсказательную способность модели в целом. С математической точки зрения метод LR предпочтительнее, чем Conditional или Wald. Кроме того, также как и линейной регрессии, из пошаговых методов предпочтительнее методы последовательного исключения. Методы пошагового ввода повышают вероятность ошибки II рода, т.е. увеличивают риск не выявить предикторы, которые на самом деле влияют на вероятность исхода (suppressor effect) [4, 10].

Давайте выполним логистический регрессионный анализ на уже знакомом нам примере Северодвинского исследования, в которое были включены 869 женщин с одноплодной беременностью и срочными родами [12-15]. Из всех имеющихся данных: возраст (переменная «vozrast»), гестационный срок (переменная «srok»), пол ребенка (переменная «pol»), а также длина (переменная «dlina») и масса тела (переменная «ves») ребенка при рождении, только пол является дихотомической переменной. Посмотрим, имеется ли какая-либо взаимосвязь между полом ребенка и его длиной, весом, гестационным возрастом, а также можно ли, и с какой точностью, определить пола ребенка, если известны перечисленные характеристики. Таким образом, зависимая переменная – пол ребенка, независимые переменные, или предикторы – длина, масса и гестационный срок ребенка.

Перед проведением анализа мы трансформируем массу тела из интервальной в порядковую переменную для того, чтобы посмотреть особенности включения в анализ порядковых переменных. В результате масса тела будет разбита на 3 категории: «nizkaya» (до 2500 гр.), «norma» (2500-3999 гр.) и «vysokaya» (4000 гр. и более). Для этого в меню «Transform» выберите «Recode into Different Variables», в результате чего откроется одноименное окно. В левом поле окна перечислены все переменные, из которых необходимо выбрать ту, которую мы хотим перекодировать. В нашем случае это «ves». Выделите ее нажатием левой кнопки мыши и перенесите в правую область, нажав на стрелку (Рис. 1).

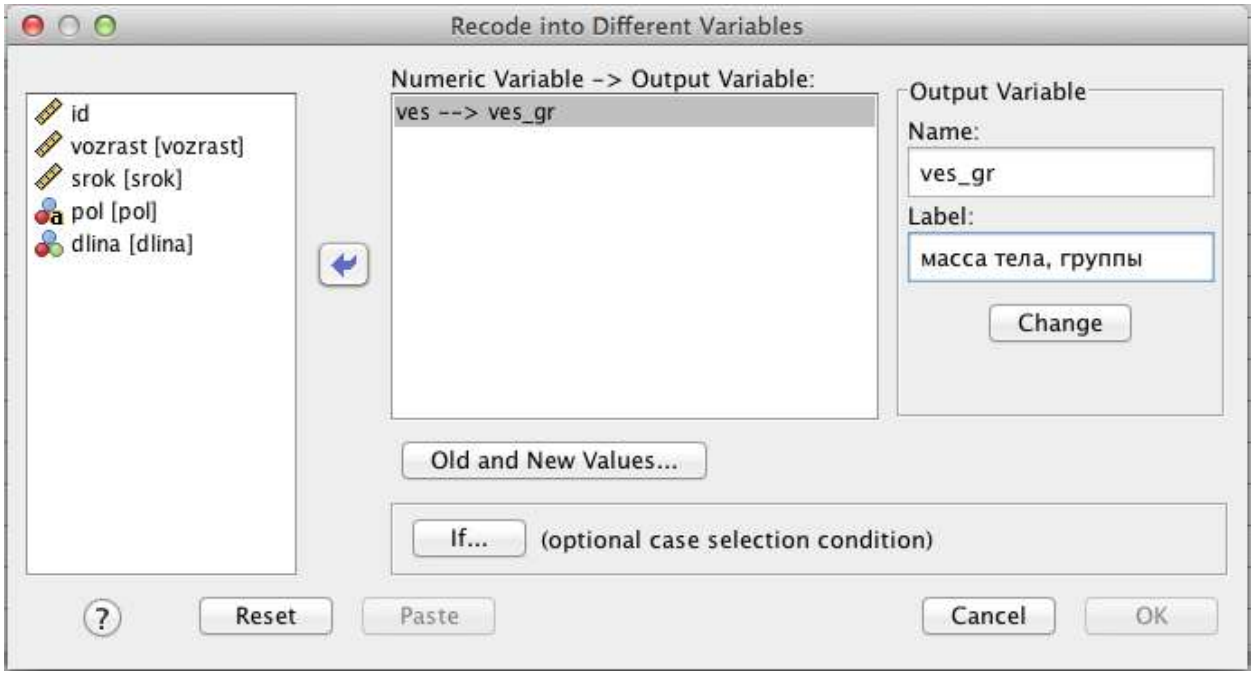


Рисунок 1. Диалоговое окно «Recode into Different Variables».

После этого в строке под названием «Name» напишите название новой переменной, например, «ves_gr», в строке «Lable» - расшифровку: «Масса тела, группы», после чего нажмите на кнопку «Change». Затем, нажатием кнопки «Old and New Values» откройте соответствующее окно (Рис. 2(A)). В левой половине окна активируйте строку «Range LOWEST through value», нажав на соответствующую точку, и введите цифру,

значения ниже которой, включая ее, войдут в категорию «nizkaya», т.е. 2499. В правой половине окна в строке «Value» введите цифру, которой эта категория будет обозначена в нашей новой переменной, например 0 (Рис. 2(A)). Затем нажмите на кнопку «Add», после чего эта категории будет добавлена в поле «Old→New». Верхняя (3999) и нижняя (2500) границы следующей категории, «normal», должны быть введены в две строки под названием «Range», новое обозначение категории «1» - в строку «Value». После нажатия на «Add» вторая категория также окажется в правом поле. Таким же образом нужно создать третью категорию «vysokaya», начиная со строки «Range value through HIGHEST». Когда все три категории будут обозначены в поле «Old→New» (Рис. 2(B)), закройте окно, нажав на «Continue», а затем и оставшееся окно кнопкой «Ok». В результате в базе будет создана новая переменная. Останется только подписать обозначения к названиям категорий (0, 1 и 2) в графе «Values» (Рис. 3).

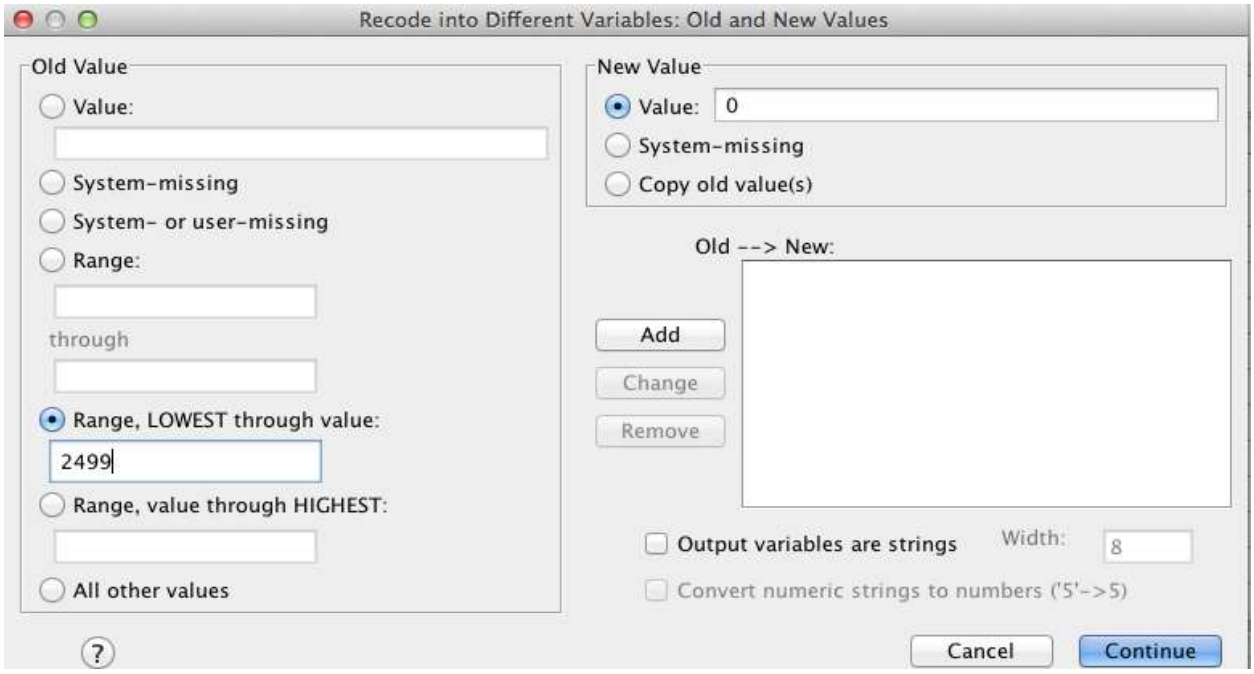


Рисунок 2(A). Диалоговое окно «Recode into Different Variables: Old and New Values».

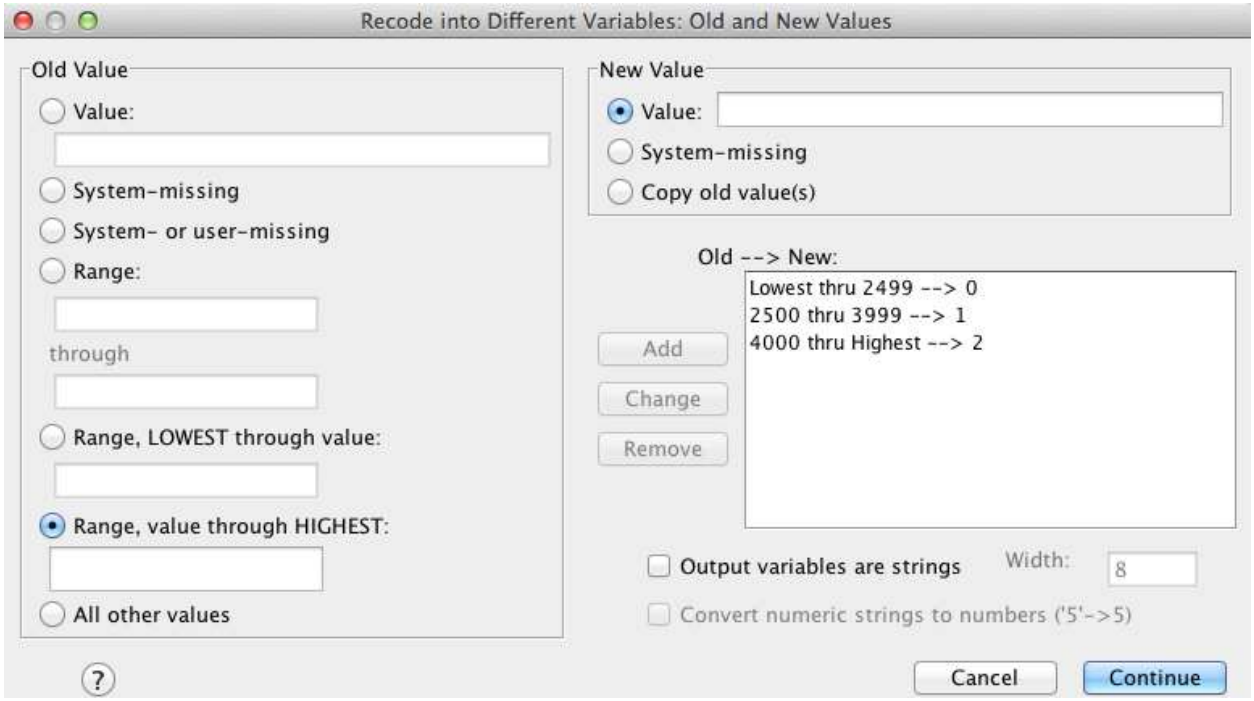


Рисунок 2(B). Диалоговое окно «Recode into Different Variables: Old and New Values».

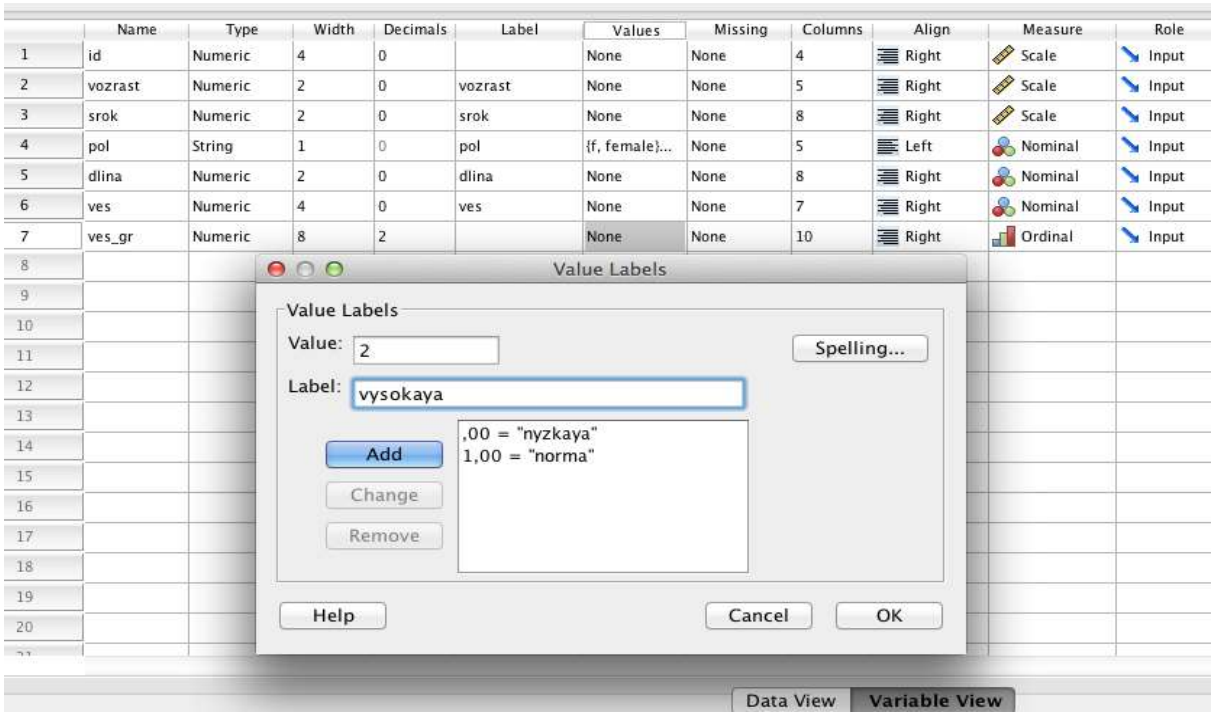


Рисунок 3. Названия категорий порядковых переменных.

В итоге переменная «ves_gr», закодирована таким образом, что наименьшая группа обозначена цифрой (0), средняя – «1», а наибольшая – «2». Также необходимо обратить внимание на то, как закодированы все остальные качественные переменные. В нашем примере это только пол: женский пол закодирован 0, а мужской – 1. Это важно для правильной интерпретации результатов в последующем, т.к. SPSS воспринимает числовые обозначения как цифры, а не как обозначения категорий.

Теперь перейдем к выполнению логистического регрессионного анализа. Логистический регрессионный анализ расположен в меню «Regression»: Analyze → Regression → Binary Logistic. Основное окно сильно напоминает таковое в линейной регрессии. Перенесите зависимую (Dependent) и независимые (Covariates) переменные в соответствующие окна, как это показано на рисунке 4. Ведем все предикторы в модель одновременно методом форсированного ввода (Enter): он используется SPSS по умолчанию, так что в графе «Method» ничего менять не нужно.

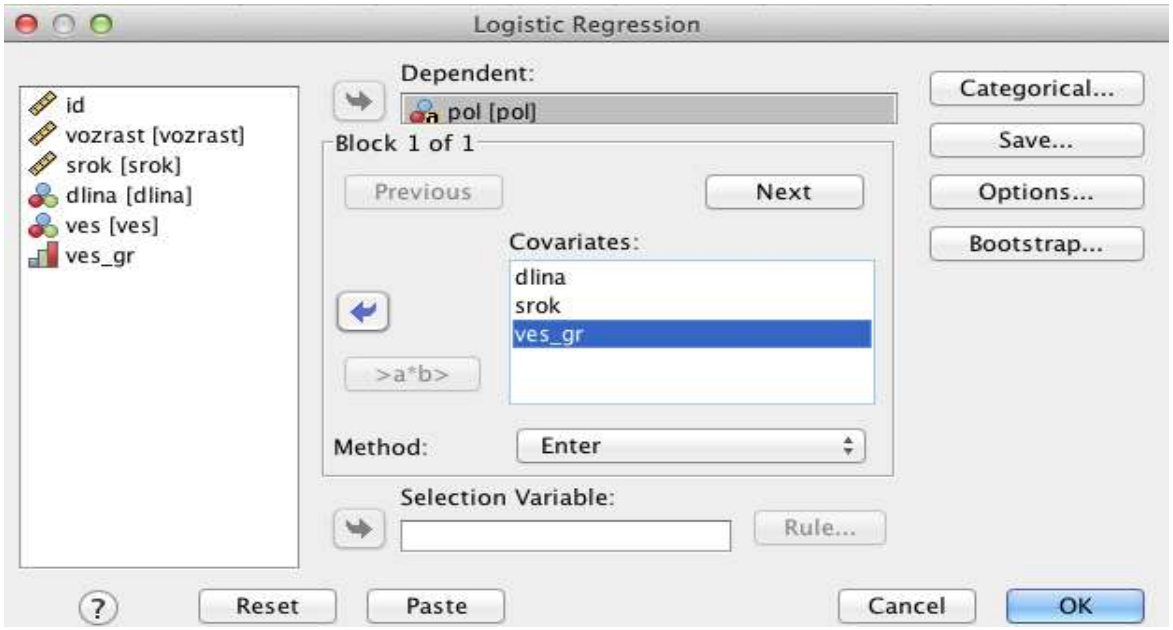


Рисунок 4. Диалоговое окно «Logistic Regression».

Как уже было сказано, SPSS воспринимает все цифровые обозначения как цифры, а все переменные как интервальные. Исходя из этого необходимо «сообщить» программе какие из переменных включаемых в модель являются категориальными. Для этого существует меню «Categorical». Нажав на кнопку с этим названием, Вы откроете окно «Logistic Regression: Define Categorical Variables» (Рис. 5). В нем из левого поля необходимо перенести в правое все категориальные переменные. В нашем случае это «ves_gr». Помимо этого нужно обозначить тип контрастирования (т.е. способ сравнения категорий признака между собой). По умолчанию SPSS использует способ «Indicator». По сути, это создание «dummy» переменных [21], которое при проведении множественного линейного регрессионного анализа мы проводили вручную. При использовании этого способа сравнения категорий остается только выбрать референс-категорию, т.е. ту, с которой все остальные будут сравниваться. Это может быть либо первая – «First», либо последняя – «Last». Если Вы хотите сравнить каждую из категорий переменной с первой (в случае с переменной «ves_gr» это была бы категория «nizkaya», обозначенная цифрой 0), то нужно активировать обозначение «First», кликнув левой кнопкой мыши на соответствующую точку. Если в качестве референс-категории Вы выбрали последнюю, то необходимо активировать «Last», что уже сделано в SPSS по умолчанию, также используем ее в анализе. После этого, нажав на клавишу «Change», Вы подтверждаете Ваш выбор референс-категории. Для того, чтобы продолжить анализ необходимо нажать на клавишу «Continue».

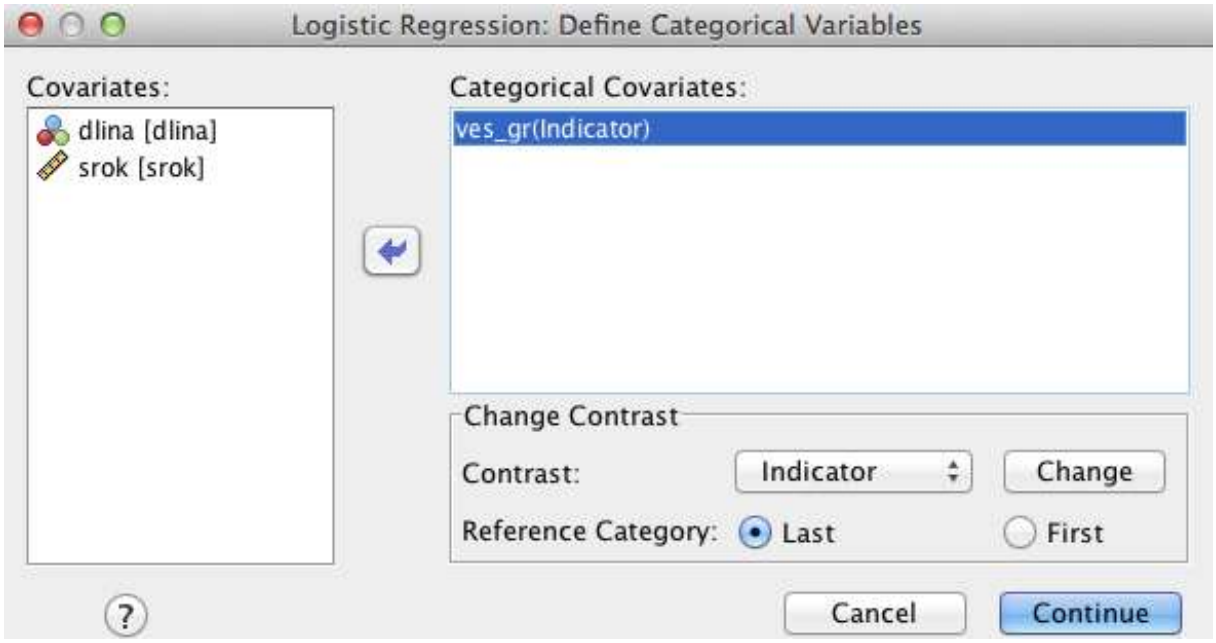


Рисунок 5. Выбор способа сравнения для категориальных переменных.

Следующее меню, которое нам необходимо, - это «Save» (рис. 6). Открыв соответствующее меню нажатием кнопки, увидим окно, схожее с таковым в линейной регрессии. Выберем Standardized residuals, Cook's distance, Leverage values, DfBeta(s), Covariance ratio/matrix, которые нам потребуются для диагностики соответствия модели имеющимся данным (подробнее рассмотрим далее). Уникальными для логистической регрессии является вычисление спрогнозированных значений вероятностей (Predicted probabilities) и предсказанной принадлежности к группе (Predicted group membership), которые будут сохранены в качестве новых переменных в файле с базой данных, что впоследствии позволит работать с ними так же, как и с имевшимися переменными. В данной версии представлена возможность внести информацию о модели в файл с расширением XML (Export model information to XML file), для того, чтобы заданные условия можно было использовать при работе с другими файлами.

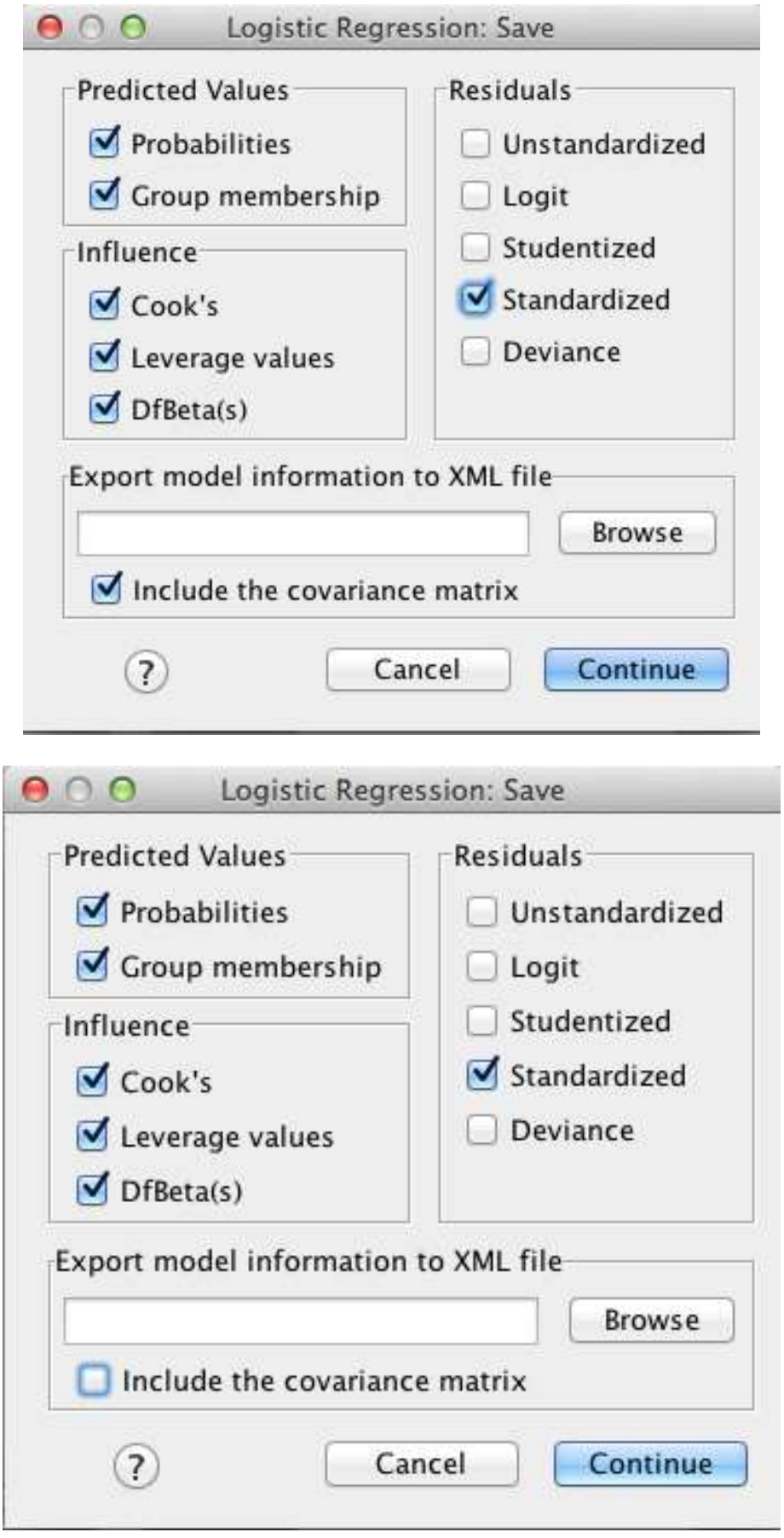


Рисунок 6. Диалоговое окно «Логистическая регрессия: Save».

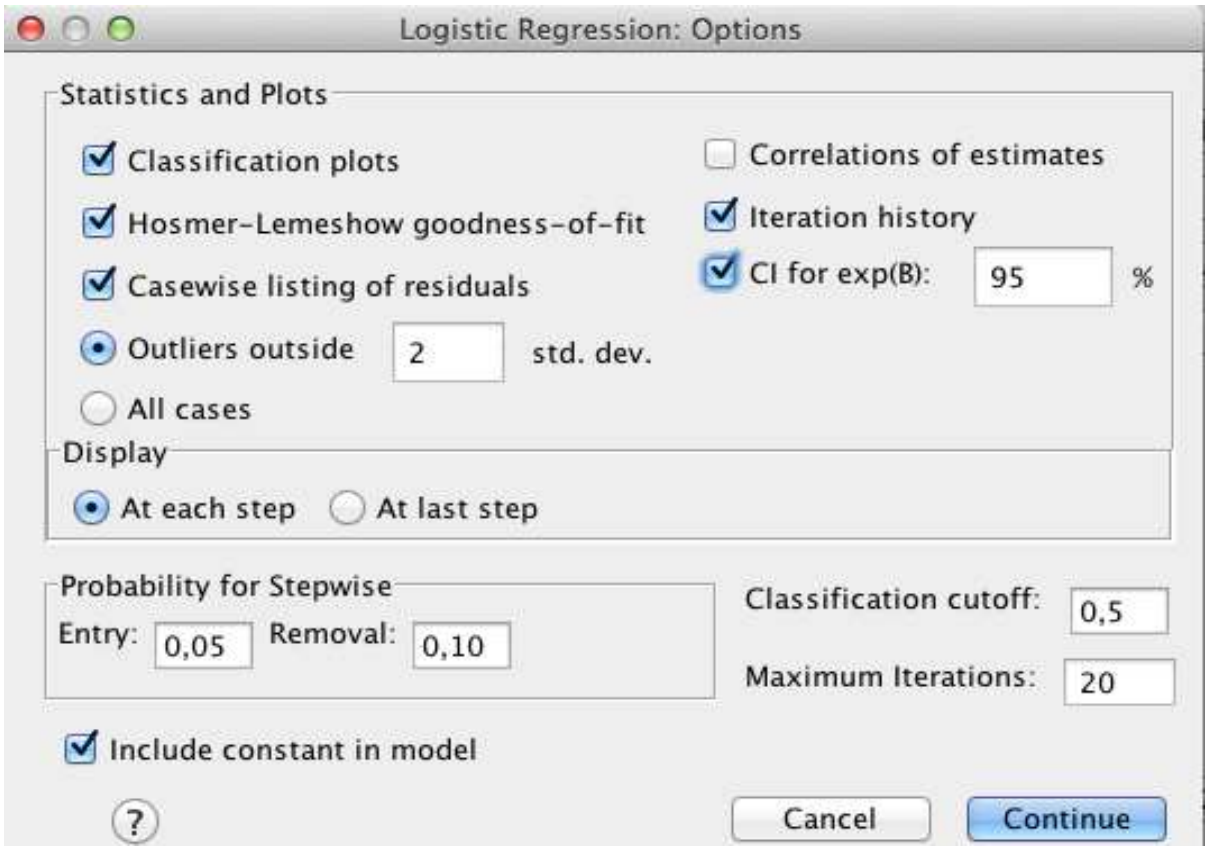


Рисунок 7. Диалоговое окно «Options».

Следующее окно, которое нас интересует, «Параметры» («Options») представлено на рис. 7. Флажок «Classification plots» позволяет включить в вывод диаграмму, в которой можно увидеть, какое значение зависимой переменной наблюдалось фактически и было предсказано с помощью построенной регрессионной модели для каждого наблюдения. Таким образом, можно будет оценить насколько адекватно построенное регрессионное уравнение отражает реальные данные. Отметим также Hosmer-Lemeshow goodness-of-fit (также показывает насколько хорошо предсказанная модель будет анализировать фактические данные), Casewise listing of residuals (формирование списка «выскакивающих» наблюдений (outliers), Iteration history (ход итераций или повторных циклов обработки информации для построения модели), CI for exp(B) (доверительный интервал для exp(B), автоматически установлен 95%). Системой также автоматически отмечено, что данные параметры надо выводить на каждом шаге построения регрессионного уравнения, что менять мы не будем, так как у нас будет всего один шаг при одномоментном вводе всех предикторов в модель. Далее указываются критерии шагового отбора данных. Автоматически для включения в модель значение вероятности должно составлять 0,05, для исключения из модели – 0,01. При желании можно указывать другие значения данные показателей, но мы оставим их без изменения. Константа представляет собой значение зависимой переменной, когда значения всех зависимых переменных равны нулю (Y-intercept). SPSS включает константу в модель автоматически, но Вы можете он нее отказаться. Следует нажать «Continue» для сохранения заданных параметров в меню «Options».

Для проведения самого логистического регрессионного анализа следует нажать на клавишу «OK», после чего автоматически откроется новое окно «Вывод» («Output»).

В первой таблице «Вывода» (рис. 8) указано, какое количество наблюдений, из имеющихся в базе данных, было включено в данный анализ. В данном примере было проанализировано 869 наблюдений. Затем представлена таблица (рис. 9) с указанием метода кодирования зависимой переменной (мужской пол (male) у нас был закодирован как «1»).

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	869	100,0
	Missing Cases	0	,0
	Total	869	100,0
Unselected Cases		0	,0
Total		869	100,0

a. If weight is in effect, see classification table for the total number of cases.

Рисунок 8. Заключение по наблюдениям, включенным в анализ.

Далее (рис. 10) мы видим таблицу кодирования категориальных переменных (в случае их отсутствия данной таблицы не будет в выводе), где автоматически программой были сформированы две «dummy» переменные, которые закодированы относительно референсной категории «Vysokaya».

Dependent Variable Encoding

Original Value	Internal Value
female	0
male	1

Рисунок 9. Кодировка зависимой переменной.

Categorical Variables Codings

		Frequency	Parameter coding (1)	Parameter coding (2)
ves_gr	nyzkaya	19	1,000	,000
	norma	772	,000	1,000
	vysokaya	78	,000	,000

Рисунок 10. Кодировка категориальных переменных.

Следует обратить внимание на то, что затем данные представлены в виде анализа в два этапа: сначала выводится анализ зависимости пола только от константы (Шаг 0), затем приводится модель, одномоментно включающая все интересующие нас предикторы (Шаг 2). Оба шага имеют однотипные таблицы представления результатов.

В первоначальной модели (Шаг 0) переменные предикторы не включаются в модель. Как видно из истории итераций (рис. 11), подбор моделей был остановлен на втором шаге, так как значения параметров изменились менее чем на 0,001.

Согласно построенной модели на основании значения константы все новорожденные будет отнесены лишь к одной категории пола. Пол будет выбран исходя из того, к какому полу в итоге относилось большинство родившихся детей в фактической базе данных. Согласно классификационной таблице (рис. 12) в базе было 443 мальчика из 869 детей, соответственно все новорожденные были отнесены к мужскому полу. Крайнее нижнее правое число в таблице указывает на процент корректно рассчитанных с помощью регрессионного уравнения значений показателя «Pol» в общей выборке. Модель правильно оценивала вероятность родиться мальчиком в 51% случаев (что несколько лучше, чем вероятность 50/50).

Далее в таблице представлены переменные, вошедшие в модель (рис. 13). Коэффициент регрессионного уравнения (B) для единственного включенного фактора

константы (b_0) составляет 0,039. Следующие столбцы в данной таблице – это стандартная ошибка коэффициента B (S.E.); критерий Вальда (Wald, критерий значимости коэффициента B для соответствующей независимой переменной; его значимость находится в прямой зависимости от самого значения критерия и от числа степеней свободы (df)); статистическая значимость по критерию Вальда (Sig., при ее значениях <0,05 введенный предиктор статистически значимо влияет на модель); Exp(B) – экспонента B или e^B, отражает изменение отношения шансов (Odds Ratio) при изменении предиктора на единицу измерения, о котором упоминалось в теоретической части статьи.

Затем следует таблица с переменными, не вошедшими в модель (рис. 14). Последняя строка (Overall Statistics) содержит информацию об остаточном значении хи-квадрат (residual chi-square) для всех не включенных факторов (27,473, статистически значимое при p<0,001), что говорит о том, что включение данных факторов в модель значительно улучшить ее предсказательную мощность. Если данное значение будет иметь статистическую значимость выше критического значения (p>0,05), это будет свидетельствовать о том, что включение в модель выбранных предикторов не повысит ее предсказательную способность, и анализ будет закончен на этом шаге. Следует отметить, что в столбце Score приводятся значения коэффициента Poa (Roa's efficient score statistic), который является аналогом коэффициента Вальда и может быть также использован, когда применение коэффициента Вальда невозможно [19]. Предиктор с наибольшим значением данного показателя на уровне значимости <0,05 будет первым включен в модель при использовании пошаговых методов ввода в модель независимых переменных.

Iteration History[a,b,c]

Iteration		-2 Log likelihood	Coefficients Constant
Step 0	1	1204,357	,039
	2	1204,357	,039

a. Constant is included in the model.

b. Initial -2 Log Likelihood: 1204,357

c. Estimation terminated at iteration number 2 because parameter estimates changed by less than ,001.

Рисунок 11. История итераций (Шаг 0).

Classification Table[a,b]

Observed			Predicted		
			pol		Percentage Correct
			female	male	
Step 0	pol	female	0	426	,0
		male	0	443	100,0
	Overall Percentage				51,0

a. Constant is included in the model.

b. The cut value is ,500

Рисунок 12. Классификационная таблица (Шаг 0).

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	,039	,068	,333	1	,564	1,040

Рисунок 13. Переменные в уравнении регрессии (Шаг 0).

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	dlina	20,643	1	,000
		ves_gr	8,376	2	,015
		ves_gr(1)	1,553	1	,213
		ves_gr(2)	3,396	1	,065
		srok	,872	1	,350
	Overall Statistics		27,473	4	,000

Рисунок 14. Переменные, не включенные в уравнение регрессии (Шаг 0).

Теперь перейдем к описанию таблиц Шага 1 (Step 1) , которые содержат информацию о модели после одномоментного ввода всех интересующих нас независимых переменных.

В таблице Истории итераций (рис. 15) мы видим, что процесс построения модели был остановлен на третьем шаге, который не принес улучшения прогностической мощности модели. Как мы уже упоминали показатель -2 Log likelihood (аналог суммы квадратов остатков в линейной регрессии) отражает какая часть информации осталась необъясненной после применения модели для нашей базы данных. Следовательно, чем меньше значение показателя, тем более адекватной является наша модель. В целом, значение -2 Log likelihood на этом этапе (1176,373) должно быть ниже, чем таковое в Шаге 0 (1204,357), что будет свидетельствовать о том, что новая модель предсказывает значения зависимой переменной более аккуратно.

Ответ на вопрос, насколько лучше стала модель в Шаге 1, представлен при оценке коэффициентов модели, это критерий хи-квадрат (аналог F-теста в линейной регрессии) (рис. 16). Хи-квадрат является критерием статистической значимости влияния всех предикторов шага, блока, модели на зависимую переменную. В связи с тем, что был использован метод форсированного ввода переменных в модель без деления на блоки (то есть у нас были один блок, один шаг и, соответственно, одна модель), мы видим, что для шага, блока и модели в целом значения показателя хи-квадрат (chi-square) одинаковы и составляют 28,028 (рассчитывается как разность между значениями -2 Log likelihood в Шаге 1 и Шаге 0: 1204,357 - 1176,373). Количество степеней свободы (df) рассчитывается, как количество предикторов в модели + 1 (константа) – количество предикторов в базовой модели (константа), то есть df=5-1=4). Как вы видите, переменная вес, распределенная на три группы, была введена в модель в виде двух «dummy» переменных, при этом категория «vysokaya» является референсной и в модели не представлена. Уровень статистической значимости <0,001, то есть данная модель предсказывает значения исхода, лучше, чем базовая. Показатель Hosmer & Lemeshow также определяет, насколько хорошо наше модель соответствует фактическим данным (рис. 18). Если мы получаем значение с уровнем значимости >0,05, то построенная модель хорошо отражает фактические данные [10].

Далее отражены показатели, рассчитывающие приближение значения R^2 (псевдо- R^2) для логистической регессионной модели (рис. 17) и отражающие долю влияния всех переменных, включенных в модель, на зависимую переменную. Значения показателей Cox & Snell, Nagelkerke и Hosmer & Lemeshow (0,032, 0,042 и 0,023) значительно отличаются друг от друга, зависят от способа их рассчета и каждый имеют ряд ограничений. Показатель

аналога R^2 Hosmer & Lemeshow рассчитан вручную, как частное от значения хи-квадрат итоговой модели (28,028), разделенного на -2 Log likelihood в Шаге 0 (1204,357) [10]. То есть, мы можем сказать, что только 2-4% вариабельности признака «пол ребенка» обусловлены введенными в модель предикторами.

Iteration History[a,b,c,d]

Iteration		-2 Log likelihood	Coefficients				
			Constant	dlina	ves_gr(1)	ves_gr(2)	srok
Step 1	1	1176,373	-3,276	,181	-,179	-,142	-,147
	2	1176,330	-3,402	,188	-,202	-,156	-,152
	3	1176,330	-3,402	,188	-,202	-,157	-,152

a. Method: Enter

b. Constant is included in the model.

c. Initial -2 Log Likelihood: 1204,357

d. Estimation terminated at iteration number 3 because parameter estimates changed by less than ,001.

Рисунок 15. История итераций (Шаг 1).

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	28,028	4	,000
	Block	28,028	4	,000
	Model	28,028	4	,000

,

Рисунок 16. Оценка коэффициентов модели.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1176,330[a]	,032	,042

a. Estimation terminated at iteration number 3 because parameter estimates changed by less than ,001.

Рисунок 17. Итоговая оценка модели (Шаг 1).

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	5,819	8	,667

Рисунок 18. Значение критерия Hosmer-Lemeshow для итоговой модели.

В классификационной таблице (рис. 19) вновь представлено сравнение прогнозируемого распределения зависимой переменной между двумя категориями. Если вероятность менее 0,5, то зависимой переменной присваивается значение 0 (принадлежность к женскому полу), если ≥ 0,5 – то 1 (к мужскому). Как мы видим, 57,0% значений было рассчитано правильно (в Шаге 0 правильно был рассчитан 51% значений), при этом модель правильно определяла вероятность рождения 57,5% девочек и 56,4% мальчиков.

Classification Table[a]

Observed			Predicted		
			pol		Percentage Correct
			female	male	
Step 1	pol	female	245	181	57,5
		male	193	250	56,4
	Overall Percentage				57,0

a. The cut value is ,500

Рисунок 19. Классификационная таблица (Шаг 1).

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	dlina	,188	,047	16,136	1	,000	1,207	1,101	1,323
	ves_gr			,290	2	,865			
	ves_gr(1)	-,202	,618	,107	1	,743	,817	,243	2,743
	ves_gr(2)	-,157	,291	,290	1	,590	,855	,483	1,512
	srok	-,152	,058	6,777	1	,009	,859	,766	,963
	Constant	-3,402	3,001	1,286	1	,257	,033		

a. Variable(s) entered on step 1: dlina, ves_gr, srok.

Рисунок 20. Переменные в уравнении регрессии (Шаг 1).

Наиболее важной для оценки результатов анализа является таблица с переменными, включенными в итоговую модель (рис. 20). Все заданные параметры в ней аналогичны таковым на рис. 13. Кроме самого параметра Exp(B), мы задали выведение в таблицу 95% ДИ для этого показателя. Границы 95% ДИ отражают, в каких пределах с 95% вероятностью находится значение коэффициента для популяции6 из которой была сформирована анализируемая выборка. Если в интервал между нижним и верхним пределом входит единица, то параметр будет статистически не значимым, что также отражает статистическая значимость (Sig.). По данной таблице можно сделать заключение, что при увеличении длины тела новорожденного на один сантиметр, шансы родиться мальчиком увеличиваются в 1,2 раза (или на 20%) при уровне значимости критерия Вальда <0,001; при увеличении срока гестации на одну неделю, шансы рождения мальчика уменьшаются в 1,16 раза (1/0,859) или на 16% (p = 0,009). Вес при рождении не оказывает значимого влияния на пол ребенка.

Correlation Matrix

		Constant	dlina	ves_gr(1)	ves_gr(2)	srok
Step 1	Constant	1,000	-,642	-,467	-,506	-,571
	dlina	-,642	1,000	,457	,476	-,257
	ves_gr(1)	-,467	,457	1,000	,570	,063
	ves_gr(2)	-,506	,476	,570	1,000	,048
	srok	-,571	-,257	,063	,048	1,000

Рисунок 21. Корреляционный матрикс.

Согласно корреляционному матриксу выявлены корреляции средней силы между длиной и весу в обеих группах, что не мешает применению модели. Только сильные корреляционные связи (>0,9) могут влиять на полученные результаты.

Диаграмма, представленная на рисунке 22, позволяет визуально оценить, насколько

хорошо построенная модель отражает реальные данные, и соответствует данным классификационной таблицы (рис. 19). По горизонтальной оси представлены предсказанные вероятности значения переменной пол, которые закодированы как первые буквы данной переменной в базе (f – female (женский), m – male (мужской)). По вертикальной оси – реальные частоты согласно базе данные, закодированные тем же образом. Каждый столбик диаграммы соответствует определенной предсказанной вероятности (от 0 до 1), а высота столбика характеризуется количеством случаев, для которых данная вероятность была предсказана моделью. Если бы модель предсказывала вероятность исхода с точностью 100%, то до значения 0,5 по оси абсцисс располагались бы только буквы f, а после него – буквы m, и чем ближе правильные значения располагались бы к 0 и 1, соответственно, тем лучше. Однако, мы видим на диаграмме, что во многих столбиках присутствуют обе буквы, и наибольшая часть значений расположена в центре диаграммы, что говорит об ошибках предсказания в соответствии с моделью.

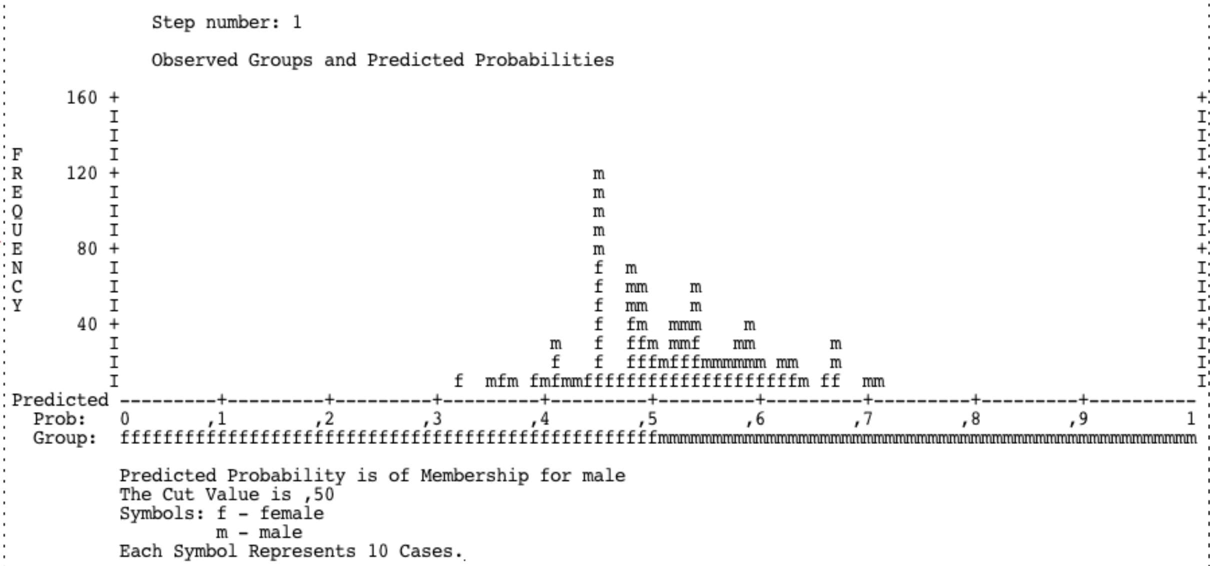


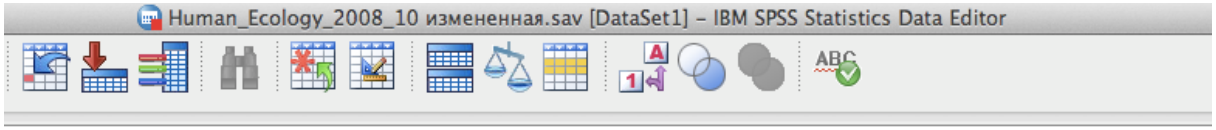
Рисунок 22. График классификации.

Перейдем к проверке необходимых условий проведения логистического регрессионного анализа.

1. зависимая переменная пол – дихотомическая (имеет лишь два варианта исхода);
2. желаемый исход в зависимой переменной (мужской пол) закодирован единицей;
3. наблюдения независимые, это не был анализ повторных измерений, оценки близнецов и т.п.;
4. отсутствие мультиколлинеарности, т.е. ситуаций, когда независимые переменные сильно коррелируют между собой (r >0,9), согласно корреляционному матриксу;
5. линейная зависимость между каждой независимой переменной и логарифмом отношения шансов (log odds);
6. независимость остатков.

Оценка остатков необходима, чтобы найти наблюдения, для которых модель работает плохо и/или которые сами оказывают влияние на модель [6, 8].

Как мы уже упоминали, уникальными для логистической регрессии являются спрогнозированные значения вероятностей (Predicted probabilities) и предсказанная принадлежность к группе (Predicted group membership). Эти остатки сохраняются как отдельные переменные в базе под названиями PRE_1 и PGR_1, соответственно (рис. 23). PRE_1 отражает предсказанную моделью вероятность отнесения каждого случая к мужскому полу. PGR_1 показывает, к какой группе отнесен каждый случай с помощью модели (то есть каждому наблюдению присвоены значения 0 или 1).



dlina	ves	ves_gr	PRE_1	PGR_1	COO_1	LEV_1	ZRE_1	DFB0_1
51	3210	1,00	,44948	0	,00350	,00285	1,10670	-,07769
50	2680	1,00	,47830	0	,00179	,00195	-,95750	-,07171
50	2975	1,00	,44059	0	,00257	,00202	1,12681	,02778
51	3372	1,00	,44948	0	,00233	,00285	-,90359	,06343
50	3376	1,00	,51626	1	,00334	,00356	,96798	,12122
55	3880	1,00	,59799	1	,01692	,01124	-1,21964	,36384
50	3080	1,00	,44059	0	,00160	,00202	-,88746	-,02188
50	3022	1,00	,55404	1	,00550	,00679	,89717	,15706
50	3070	1,00	,36757	0	,01185	,00684	1,31170	-,09590
51	3350	1,00	,44948	0	,00233	,00285	-,90359	,06343
51	3580	1,00	,52525	1	,00152	,00168	,95071	,02826
52	3770	1,00	,53423	1	,00166	,00190	,93374	-,06106
52	3650	1,00	,53423	1	,00218	,00190	-1,07096	,07003
52	4110	2,00	,60961	1	,01109	,01703	,80025	,16341
50	3100	1,00	,44059	0	,00257	,00202	1,12681	,02778
50	3230	1,00	,47830	0	,00213	,00195	1,04438	,07821
53	3080	1,00	,54318	1	,00354	,00419	,91707	-,14731

Рисунок 23. Вид в базе данных рассчитанных остатков.

Способы оценки наличия «выскакивающих» случаев (outliers): 1 - визуальный (на скаттерограмме), может быть затруднен при построении множественной регрессионной модели; 2 – оценить количество наблюдений, стандартизованные остатки которых 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%) стандартных отклонения. Стандартизованные остатки при их отметке в окне «Save» были сохранены в базе под названием «ZRE». В нашем примере таких случаев найдено не было (мы отмечали в окне «Options» флажок «Casewise listing of residuals» (рис. 7).

Для выявления единичных случаев, сильно влияющих на модель (influential cases), мы также используем такие показатели, как дистанция Кука (Cook's distance), показатель DFBeta (любое значение более единицы свидетельствует о влиянии случая на модель), показатель «рычаг» (leverage, определяется формулой (количество предикторов+1)/объем выборки, его значение находится в пределах от нуля (показатель не оказывает никакого влияния на модель) до единицы (показатель абсолютно влияет на модель) [8, 16]. Показатели сохраняется в базе под названием «COO», «DFB» и «LEV», соответственно (рис. 23).

Более подробно о способах выявления outliers и influential cases нами было написано в предыдущей статье [4]. Выявление данных случаев в логистической регресии полезно для более подробного изучения этих случаев и нахождения потенциальных ошибок при внесении информации в базу данных. В отличии от линейного регрессионного анализа в логистическом нельзя исключить данные случаи из анализа для улучшения модели.

Один из возможных вариантов представления результатов проведенного множественного логистического регрессионного анализа представлен в таблице 1. По желанию, в таблицу можно включать или не включать предикторы, которые были не значимыми. Однако, по нашему мнению все анализируемые показатели следует включить в таблицу, чтобы у читателя сложилась полная картина того, что было изучено. Следует также включить значение константы, чтобы при желании была возможность построения регрессионной модели. Таким образом, информационная способность модели с целью прогнозирования пола ребенка при рождении

составляет 57,0% (p<0,001), и наша модель хорошо соответствует фактическим данным. Значимыми факторами для определения пола ребенка являются длина при рождении и срок беременности, при этом увеличение длины повышает вероятность рождения мальчика, а увеличение срока снижает эту вероятность.

Таблица 1.

Результаты множественного логистического регрессионного анализа.

Показатель	B (SE)	95% ДИ для exp B		
		Lower	Exp B	Upper
Константа	-3,40 (3,00)		0,033	
Длина	0,19 (0,05)**	1,1	1,21	1,3
Срок	-0,15 (0,06)*	0,77	0,87	0,96
Вес <2499 г	-0,2 (0,62)	0,24	0,82	2,74
Вес 2500-3999 г	-0,15 (0,29)	0,77	0,86	0,96

Примечание. R^2 = 0,02 (Hosmer & Lemeshow), 0,032 (Cox & Snell), 0,042 (Nagelkerke).
Хи-квадрат модели 28.03, p<0.001. * <0,01, ** <0,001

Нередко логистические модели используются для диагностики каких-либо состояний или исходов. Для нашего примера это не очень показательно, однако мы опишем характеристики диагностической модели для их общего понимания и возможного дальнейшего практического применения. Если бы нас интересовал именно вариант рождения мальчика, и в будущем мы планировали бы использовать данную модель для диагностики рождения мальчиков, то модель мы бы дополнительно оценивали с помощью следующих показателей:

Чувствительность – процентное выражение частоты только истинно положительных результатов (значений исхода, равных 1) согласно модели относительно всех исходов равных 1, то есть относительная частота распределения мальчика в группу мальчиков.

Специфичность – процентное выражение частоты только истинно отрицательных результатов (значений исхода, равных 0) согласно модели относительно всех исходов равных 0, то есть относительная частота распределения девочки в группу девочек.

Безошибочность / Точность – относительная частота принятия безошибочных распределений (как в группу мальчиков, так и девочек).

Ложноотрицательный ответ (α ошибка или ошибка первого рода) – относительная частота распределения мальчика в группу девочек.

Ложноположительный ответ (β ошибка или ошибка второго рода) – относительная частота распределения девочек в группу мальчиков.

Таблица 2.

Таблица фактических и прогнозируемых частот распределения детей по полу при рождении.

Фактические значения	Прогнозируемые				Всего наблюдений	
	Мальчики		Девочки			
Мальчики	a	245	b	181	a+b	426
Девочки	c	193	d	250	c+d	443
Всего	a+c	438	b+d	431	a+b+c+d	869

Если присвоить определенные значения данным классификационной таблицы с рис. 19, как указано в таблице 2, то для определения вышеописанных показателей можно использовать следующие формулы [5]:

- чувствительность = 100*a/(a+b) =100*245/(245+181) = 57,51%
- специфичность = 100*d/(c+d) =100*250/(193+250) = 56,43%
- безошибочность/точность = 100*(a+d)/(a+b+c+d) = 100*(245+250)/(245+181+193+250) = 56,96%
- ложноотрицательный ответ = 1 – чувствительность = 100*b/(a+b) = 42,49%
- ложноположительный ответ = 1 – специфичность = 100*c/(c+d) = 43,57%

На основании проведенных расчетов можно сказать, что наша модель обладает невысокой чувствительностью (57,5%) и специфичностью (56,4%) для диагностики рождения мальчиков, поэтому ее не рекомендуется использовать на практике.

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. *Наследов А.* SPSS 19: профессиональный статистический анализ данных. СПб.: Питер, 2011. 400 с.

4. *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.

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

6. *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.

7. *Bewick V., Cheek L., Ball J.* Statistics review 14: Logistic regression, Crit Care. 2005; 9(1): 112–118.

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

9. *Cox D.D., Snell E.J.* The Analysis of Binary Data (2nd ed.). London: Chapman and Hall, 1989. 247 p.

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

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

12. *Grjibovski A., Bygren L.O., Svartbo B.* Socio-demographic determinants of poor infant outcome in north-west Russia // Paediatr Perinat Epidemiol. 2002. N 3. P. 255-62.

13. *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. N 12. P. 1159-66.

14. *Grjibovski A. M., Bygren L.O., Svartbo B., Magnus P.* Social variations in fetal growth in Northwest Russia: an analysis of medical records // Ann of Epidemiol. 2003. N 9. P. 599-605.

15. *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. N 6. P. 757-63.

16. *Hoaglin D.C., Welsch R.E.* The Hat Matrix in Regression and ANOVA // The American statistician. 1978. N 1. P. 17–22.

17. *Menard S.* Applied logistic regression analysis (2nd ed.). London: SAGE Publications Ltd., 2001. 128 p.

18. *Nagelkerke N.D.* A note on a general definition of the coefficient of determination // Biometrica. 1991. N 78. P. 691-692.

19. *Rao C.R.* In advances in ranking and selection, multiple comparisons and reliability. Birkhauser, 2005. P. 3-20.

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

21. *Suits D.B.* Use of Dummy Variables in Regression Equations // Journal of the American Statistical Association. 1957. N 280. P. 548–551.

References:

1. Grjibovski A.M. Odnofactornyj lineinyj regressionnyj analiz. [Simple linear regression analysis*]. Ekologiya cheloveka* [Human ecology (Russian Federation)] 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. Nasledov A. *SPSS 19: professional'nyi statisticheskii analiz dannykh* [SPSS 19:

professional statistical data analysis]. - SPb.: Piter, 2011. - 400 p. [in Russian].

4. Sharashova E.E., Kholmatova K.K., Gorbatova M.A., Grjibovski A.M. Primenenie mnozhestvennogo lineinogo regressionnogo analiza v zdravoohranenii. [The application of multiple logistic regression analysis in health sciences using SPSS software]. *Nauka i zdravookhranenie* [Science & Health Care] 2017. №3. С. 5-31. [in Russian]

5. 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].

6. 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.

7. Bewick V., Cheek L., Ball J. *Statistics review 14: Logistic regression*, Crit Care. 2005; 9(1): 112–118.

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

9. Cox D.D., Snell E.J. *The Analysis of Binary Data (2nd ed.).* London: Chapman and Hall, 1989. 247 p.

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

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

12. Grjibovski A., Bygren L.O., Svartbo B. Socio-demographic determinants of poor infant outcome in north-west Russia. *Paediatr Perinat Epidemiol.* 2002. N 3. P. 255-62.

13. 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. N 12. P. 1159-66.

14. Grjibovski A. M., Bygren L.O., Svartbo B., Magnus P. Social variations in fetal growth in Northwest Russia: an analysis of medical records. *Ann of Epidemiol.* 2003. N 9. P. 599-605.

15. 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. N 6. P. 757-63.

16. Hoaglin D.C., Welsch R.E. The Hat Matrix in Regression and ANOVA. *The American statistician.* 1978. N 1. P. 17–22.

17. Menard S. *Applied logistic regression analysis (2nd ed.).* London: SAGE Publications Ltd., 2001. 128 p.

18. Nagelkerke N.D. A note on a general definition of the coefficient of determination. *Biometrica.* 1991. N 78. P. 691-692.

19. Rao C.R. *In advances in ranking and selection, multiple comparisons and reliability.* Birkhauser, 2005. P. 3-20.

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

21. Suits D.B. Use of Dummy Variables in Regression Equations. *Journal of the American Statistical Association*. 1957. N 280. P. 548–551.

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; 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).