УДК 519.233.5

ОДНОФАКТОРНЫЙ ЛИНЕЙНЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ

© 2008 г. **А. М. Гржибовский**

Национальный институт общественного здоровья, г. Осло, Норвегия

В статье рассматривается применение линейного регрессионного анализа для ситуаций с одной зависимой и одной независимой переменной с использованием пакета статистических программ SPSS. Особое внимание уделяется проверке соблюдения необходимых условий для применения линейного регрессионного анализа. Изложенный материал дает общие сведения о возможности расчета значений одной переменной на основании значений другой и призван вызвать интерес читателей журнала к прочтению специализированной литературы перед началом работы над будущими публикациями.

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

В предыдущем выпуске жернала мы рассмотрели применение корреляционного анализа для определения силы и направления связи между переменными. Регрессионный анализ помимо определения силы и направления связи между переменными дает возможность прогнозировать (предсказывать) значения зависимой переменной (переменной отклика) по известным значениям независимой переменной (предиктора). В данной статье будет рассматриваться простейший вид регрессионного анализа (simple linear regression) с двумя переменными (одна зависимая и одна независимая), между которыми подразумевается линейная зависимость, которую можно в общем виде представить в виде уравнения

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon,$$

где Y – значение зависимой переменной, X – значение переменной-предиктора, β_0 – константа, β_1 – коэффициент регрессии, а ε – случайная ошибка модели. При проведении исследования на определенной выборочной совокупности измеряются значения переменных Y и X, на основании которых с помощью метода наименьших квадратов (least squares method) рассчитываются константа (constant, b_0) и коэффициент регрессии (coefficient, b_1), которые являются выборочными оценками генеральных параметров β_0 и β_1. Таким образом, если представленное выше уравнение простой линейной регрессии справедливо для генеральной совокупности, то для выборки уравнение будет иметь вид

$$\overline{\overline{Y}} = b_0 + b_1 \cdot X + \varepsilon,$$

где $\overline{\overline{Y}}$ обозначает предсказанные значения переменной Y, а X – фактические значения переменной X, полученные в ходе исследования. Каким образом рассчитываются константа и коэффициент регрессии с помощью метода наименьших квадратов? Как построить простую регрессионную модель в SPSS? Как интерпретировать результат? Чтобы ответить на эти вопросы, рассмотрим взаимосвязь между длиной и массой тела новорожденных из Северодвинского когортного исследования [2]. Для примера отобраны только дети первородящих женщин, рожденные в срок от одноплодных беременностей. Файл «Human_Ecology_2008_10.sav» можно скачать с сайта журнала «Экология человека» : www.nsmu.ru/nauka_sgmu/rio/eco_human. Переменные «id», «vozrast», «srok», «pol», «dlina», «ves» обозначают идентификационный номер участниц исследования, возраст (полных лет), гестационный возраст, пол ребенка, длину и массу тела ребенка при рождении соответственно.

Прежде чем применять линейный регрессионный анализ, следует убедиться, что исследуемая взаимосвязь носит линейный характер. В про-

тивном случае метод линейной регрессии неприменим, а прогнозирование значений переменной Y на основании вышеприведенного уравнения будет ошибочным. Для проверки достаточно построить скаттерограмму (график разброса), используя меню «Scatter/Dot» (рис. 1), которое можно найти в выпадающем меню «Graphs».

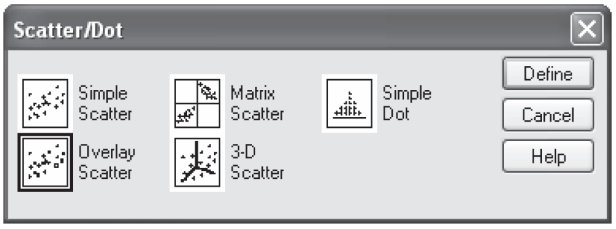

Рис. 1. Окно «Scatter/Dot»

Для построения простой скаттерограммы следует выбрать «Simple Scatter», после чего появится окно «Simple Scatterplot» (рис. 2), в котором также можно переместить интересующие нас переменные из левого поля в одно из правых в зависимости от поставленной задачи. На рис. 2 показано, как выбрать переменные для построения скаттерограммы с длиной тела новорожденных, отложенной на оси абсцисс, и весом – на оси ординат. Построение скаттерограммы запускается нажатием на кнопку «OK».

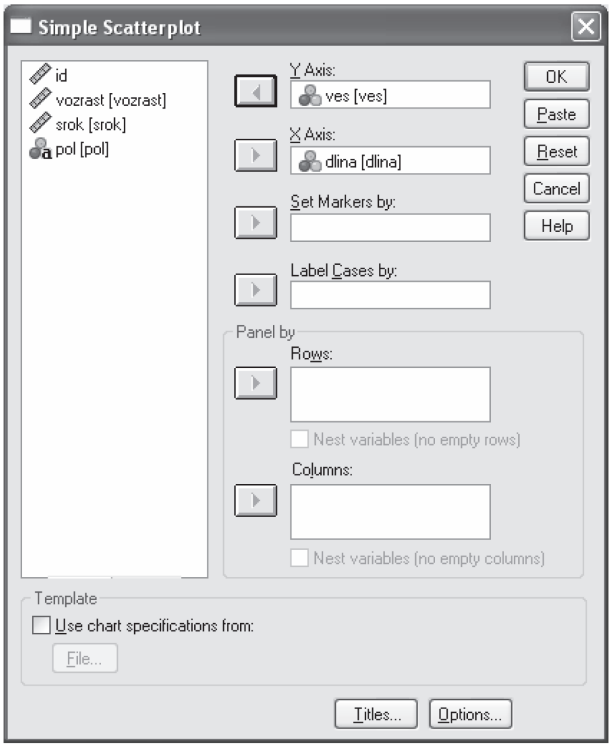

Рис. 2. Окно «Simple Scatterplot»

Полученная скаттерограмма свидетельствует о том, что зависимость между длиной и массой тела новорожденных носит линейных характер (рис. 3). Двойной клик левой кнопкой мыши в поле скаттерограммы вызовет меню «Chart Editor», в котором можно запросить у SPSS построить прямую по методу наименьших квадратов путем нажатия на кнопку, обведенную кружком, на рис. 4, после чего появится меню «Properties», показанное на рис. 5. В нем по умолчанию уже установлено построение прямой, хотя возможно построение кривых, заданных квадратным (Quadratic), кубическим (Cubic) уравнениями. Также возможно построение доверительных интервалов (Confidence Intervals) для средних предсказанных значений зависимой переменной (Mean), а также интервалов для индивидуальных наблюдений (Individual).

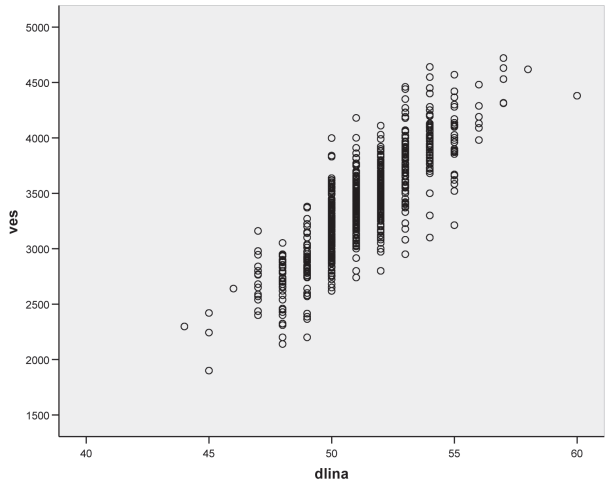

Рис. 3. Скаттерограмма для длины и массы тела новорожденных в г. Северодвинске

Из рис. 4. видно, что прямая хорошо соответствует данным, но вокруг нее существует определенный разброс, то есть прямая не идеально предсказывает значение массы тела при известной длине. Однако метод наименьших квадратов обеспечивает то, что сумма квадратов разностей между фактическими значениями переменной Y от предсказанных уравнением регрессии значений $\overline{\overline{Y}}$ будет минимальна. Разность между предсказанным значением $\overline{\overline{Y}}$ и фактическим значением Y, полученным в результате исследования, называется остатком. Графически остатки представляют собой отрезки, отложенные параллельно оси ординат из каждой точки скаттерограммы на регрессионную прямую (рис. 6). Поскольку объем выборки составляет 869 человек, на рис. 6 отложено 869 остатков на регрессионную прямую. Рассмотрим одно наблюдение, при котором длина тела новорожденного равна 60 см (крайняя правая точка на скаттерограмме). Для него остатком будет отрезок от точки, обозначенной Y, до точки, обозначенной $\overline{\overline{Y}}$. Кроме того, на рис. 6 проведена прямая, показывающая среднее значение зависимой переменной ($\overline{Y}$). Если между зависимой и независимой переменными нет никакой связи, то регрессионная прямая будет проходить параллельно оси абсцисс через среднее значение зависимой переменной, то есть среднее значение зависимой переменной будет одинаково независимо от значения независимой переменной.

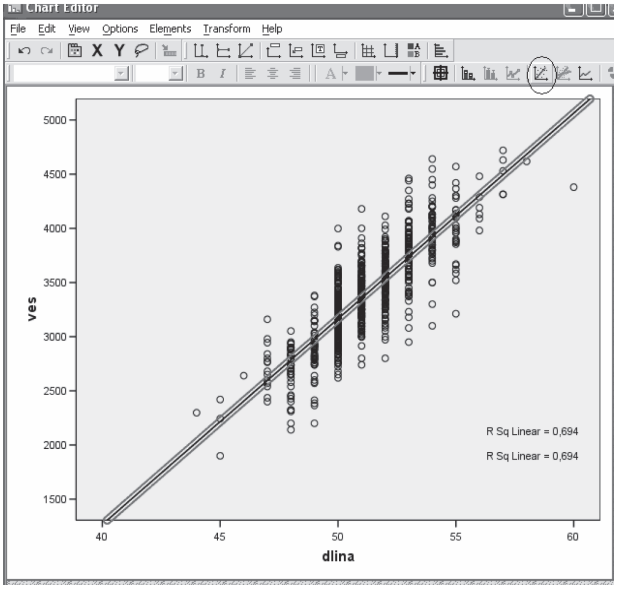

Рис. 4. Скаттерограмма взаимосвязи длины и массы тела новорожденных в г. Северодвинске с регрессионной прямой, построенной SPSS автоматически с использованием метода наименьших квадратов

Сумма квадратов наряду с дисперсией и стандартным отклонением является мерой рассеяния вокруг среднего. Общая сумма квадратов (total sum of squares, SS_t) для всех значений зависимой переменной является мерой общей вариации переменной Y:

$$SS_t = \sum_{i=1}^{n} \left(\bar{Y}_i - \bar{Y} \right)^2.$$

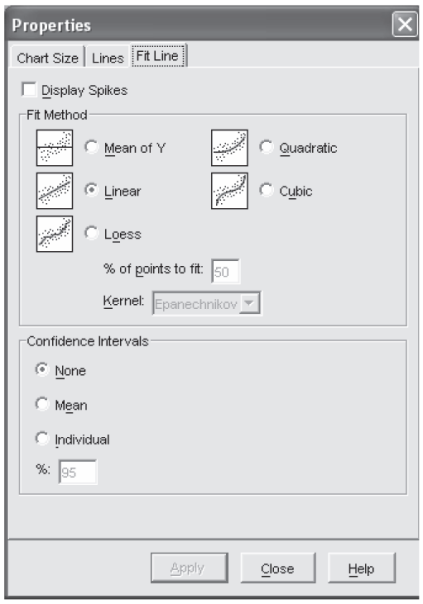

Рис. 5. Окно «Properties»

Общая вариация зависимой переменной может быть представлена в виде суммы вариации, которая может быть объяснена с помощью модели (modell sum of squares, SS_m), и остаточной вариации (residual sum of squares, SS_r), которую модель не объясняет.

$$SS_t = SS_m + SS_r.$$

Поскольку остатки представляют собой отклонения фактических значений от предсказанных, то сумма квадратов остатков будет являться мерой неточности построенной модели, а точнее, мерой вариации переменной Y, которую невозможно объяснить моделью (модель строится так, чтобы сумма именно этих квадратов была минимальна):

$$SS_r = \sum_{i=1}^{n} \left(Y_i - \bar{\bar{Y}}_i \right)^2.$$

Отсюда следует, что чем больше значение остаточной вариации, тем меньшую долю вариации зависимой переменной способна объяснить модель. Вариацию, обусловленную моделью, можно представить либо в виде $SS_m = SS_t - SS_r$, либо в виде

$$SS_m = \sum_{i=1}^{n} \left(\bar{\bar{Y}}_i - \bar{Y} \right)^2.$$

Доля общей вариации, которую способна объяснить регрессионная модель, выражается в виде уже знакомого нам из предыдущей статьи коэффициента детерминации (R^2).

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

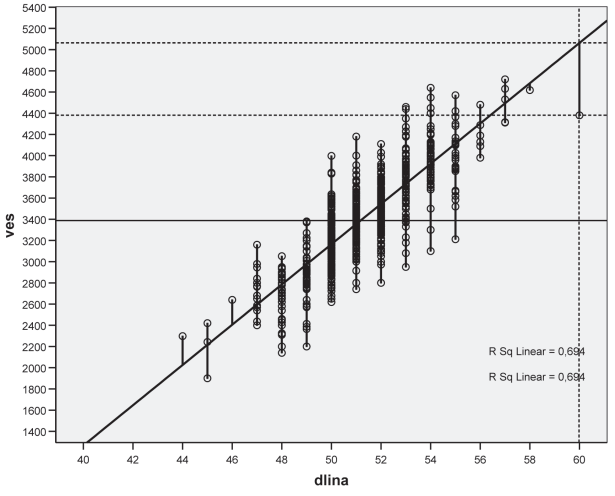

Рис. 6. Скаттерограмма с регрессионной прямой, остатками и средним значением зависимой переменной

Все суммы квадратов и коэффициент детерминации рассчитываются SPSS автоматически. Кроме того, уже при построении регрессионной прямой (см. рис. 4) SPSS выдает значение коэффициента детерминации (R Sq Linear = 0,694), который показывает, что 69 % вариабельности массы тела новорожденных может быть объяснена различиями в длине. Кроме того, SPSS автоматически проверяет статистическую значимость построенной линейной модели с помощью критерия F, уже известного нам из дисперсионного

анализа, а также представляет значения константы и коэффициента, которые можно будет использовать для прогнозирования зависимой переменной по известным значениям переменной-предиктора.

Для проведения линейного регрессионного анализа необходимо на панели инструментов выбрать выпадающее меню «Analyze», в которым следует выбрать меню «Regression», а затем «Linear», в результате чего появится диалоговое окно «Linear Regression» (рис. 7), в котором необходимо выбрать и поместить в соответствующие поля зависимую (Dependent) и независимую (Independent) переменные. Для данного примера зависимой переменной является «ves», а независимой «dlina».

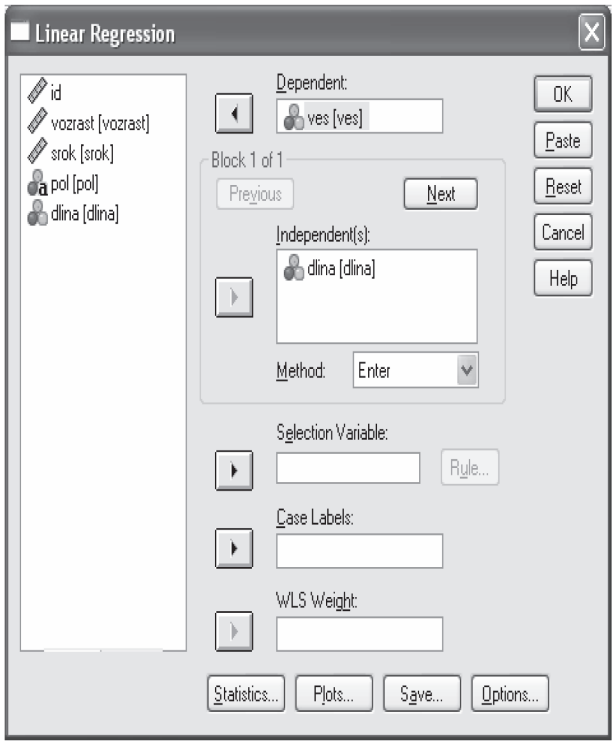

Рис. 7. Диалоговое окно «Linear Regression»

После определения зависимой и независимой переменных следует в меню «Statistics» (рис. 8) отметить (помимо Estimates и Model Fit, отмеченных по умолчанию) Confidence Intervals (доверительные интервалы для константы и коэффициента), Descriptives (описательная статистика для обеих переменных), Durbin-Watson (критерий Дарбина – Уотсона, который проверяет соблюдение условия независимости наблюдений), а также Casewise diagnostics для определения атипичных наблюдений. SPSS по умолчанию считает атипичными наблюдения, для которых остатки превышают 3 стандартных отклонения от предсказанных величин. Рекомендуется поменять это число на 2, как показано на рис. 8. Возвращение в диалоговое окно «Linear regression» осуществляется нажатием на кнопку «Continue».

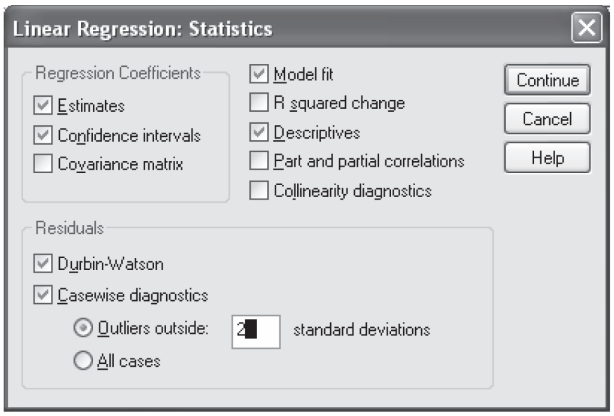

Рис. 8. Диалоговое окно «Linear Regression: Statistics»

В диалоговом окне «Linear Regression» (см. рис. 7) следует также войти в меню «Plots», в котором следует отметить все Standardized Residual Plots, а также запросить график зависимости стандартизованных остатков от стандартизованных предсказанных значений, как показано на рис. 9. Что каждый из графиков обозначает и зачем их необходимо строить, будет объяснено чуть позже. Возвращение в диалоговое окно «Linear regression» осуществляется нажатием на кнопку «Continue»

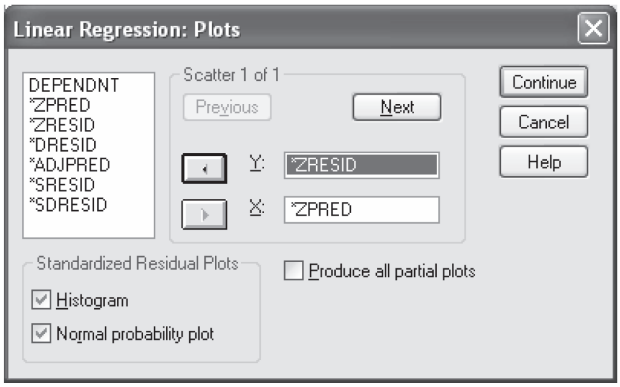

Рис. 9. Диалоговое окно «Linear Regression: Plots»

Помимо этого в окне «Linear Regression» (см. рис. 7) можно открыть окно «Save» для того, чтобы отметить, какую дополнительную информацию можно сохранить в виде новых переменных при проведении регрессионного анализа, а также для диагностики модели. Учитывая, что мы рассматриваем сейчас простейшую модель, можно отметить только Unstandardized predicted values (предсказанные значения, или Y), Unstandardized residuals (абсолютные значения остатков, или $Y - \overline{Y}$), а также оба из предлагаемых 95 % интервалов (для среднего значения зависимой переменной для каждого значения X и для индивидуальных значений зависимой переменной для каждого значения X), как показано на рис. 10. Остальные опции более полезны для многомерного линейного регрессионного анализа, когда одна зависимая переменная прогнозируется с помощью нескольких независимых переменных. Возвращение в диалоговое окно «Linear regression» осуществляется нажатием на кнопку «Continue». Запуск анализа производится нажатием на «OK».

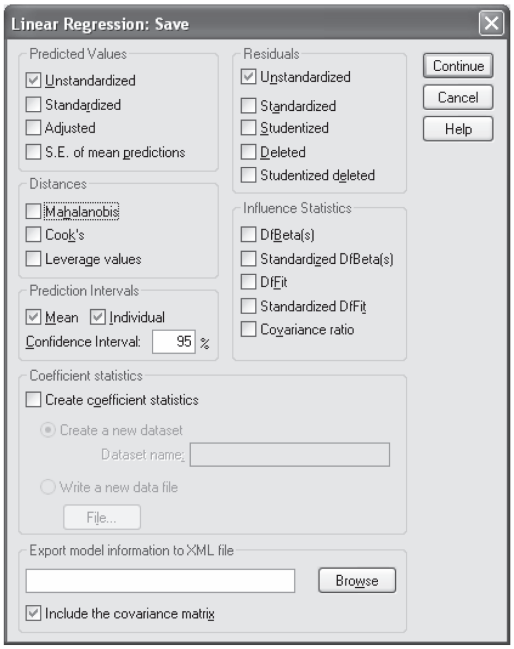

Рис. 10. Диалоговое окно «Linear Regression: Save»

Первая таблица результатов представляет собой описательную статистику для обеих переменных (рис. 11). Рассчитываются средние арифметические (Mean), стандартное отклонение (Std. Deviation) и указывается число наблюдений по каждой из переменных (N). Поскольку пропущенных значений в данном файле нет – эти значения одинаковы. При наличии пропущенных значений в анализ будут включены только те участники исследования, для кого имеются данные по обеим переменным.

Descriptive Statistics

	Mean	Std. Deviation	N
ves	3388,20	435,806	869
dlina	51,18	1,913	869

Рис. 11. Описательная статистика

Следующая таблица результатов (рис. 12) предствляет коэффициент корреляции Пирсона для оценки связи между переменными (Pearson Correlation) с указанием уровня значимости для одностороннего теста (Sig, хотя нужно представлять результаты двусторонних тестов), а также число пар наблюдений (N). Видно, что между переменными имеется сильная прямая корреляционная связь ($r = 0,83$; $p < 0,001$, $n = 869$).

Correlations

		ves	dlina
Pearson Correlation	ves	1,000	,833
	dlina	,833	1,000
Sig. (1-tailed)	ves	.	,000
	dlina	,000	.
N	ves	869	869
	dlina	869	869

Рис. 12. Результаты корреляционного анализа

В следующей таблице представлены общие сведения о модели (рис. 13), из которых самым нужным является коэффициент детерминации (R Square), равный 0,694, то есть более 69 % вариабельности массы тела новорожденных может быть обусловлено различиями в длине. Критерий Durban-Watson предназначен для проверки соблюдения условия независимости наблюдений. В идеальной ситуации он равен 2,0. Допустимые значения – от 1 до 3. Если данный критерий имеет значение менее 1 или более 3, это означает, что условие независимости остатков не соблюдается, а значит, прогнозирование с помощью этого метода будет не совсем корректным. К слову, при прогнозировании частоты встречаемости заболеваний в зависимости от календарного года это условие практически никогда не выполняется! В данном случае проблем с зависимостью нет, так как значение критерия близко к 2.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,833[a]	,694	,694	241,115	2,039

a. Predictors: (Constant), dlina

b. Dependent Variable: ves

Рис.13. Общие данные о модели

В следующей таблице результатов (рис. 14), ANOVA) представлены результаты применения критерия F, с помощью которого проводится проверка значимости модели. В таблице также представлены общая сумма квадратов (Total Sum of Squares, SS_t), сумма квадратов остатков (Residual Sum of Squares, SS_r) и сумма квадратов модели (Regression Sum of Squares, SS_m). Помня определение коэффициента детерминации, нетрудно его рассчитать из представленных сумм квадратов ($1,1*10^8/1,6*10^8$). Результат будет немного отличаться от коэффициента детерминации из рис. 13 по причине округления. Абсолютное значение критерия F представляет собой отношение средних квадратов модели и средних квадратов остатков. Средние квадраты рассчитываются путем деления суммы квадратов (столбец Sum of Squares) на число степеней свободы (столбец df). Для данного примера $F = 1\,969$, что соответствует достигнутому уровню значимости (столбец Sig.), $p < 0,001$, и говорит о том, что нулевую гипотезу об отсутствии взаимосвязи между переменными можно отвергнуть.

ANOVA[b]

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			

a. Predictors: (Constant), dlina

b. Dependent Variable: ves

Рис. 13. Результат применения критерия F для определения значимости регрессионной модели

В следующей таблице представлена самая важная информация, а именно значения константы и коэффициента с указанием уровней значимости и доверительных

интервалов. Гипотеза о равенстве коэффициентов нулю проверяется с помощью критерия Стьюдента автоматически (рис. 14). Если коэффициент равен нулю, то регрессионная прямая будет параллельна оси абсцисс (угол наклона равен нулю), то есть значение зависимой переменной будет постоянно независимо от значений переменной-предиктора.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
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

a. Dependent Variable: ves

Рис. 14. Таблица коэффициентов регрессии

Значения константы (b_0) и коэффициента регрессии (b_1) с их стандартными ошибками представлены в столбце Unstandardized Coefficients. Если подставить полученные значения в уравнение, то итоговое уравнение регрессии для взаимосвязи длины и массы тела новорожденных можно представить в виде

$$\overline{\overline{Y}} = -6327 + 190X.$$

Значение коэффициента показывает, насколько увеличится значение зависимой переменной при увеличении независимой переменной на единицу. Для данного примера масса тела новорожденных увеличивается на 190 г при увеличении длины на 1 см. Стандартизованный коэффициент (Standardized Coefficient) показывает, на сколько стандартных отклонений уваличится зависимая переменная при увеличении независимой переменной на одно стандартное отклонение. На практике стандартизованные коэффициенты используются нечасто. Кроме того, SPSS рассчитывает 95 % доверительные интервалы (Confidence intervals) для коэффициента и константы. В данном случае если наша выборка отражает структуру генеральной совокупности, то коэффициент для генеральной совокупности будет с 95 % надежностью находиться в пределах от 181 до 192. Константа показывает, где регрессионная прямая пересечет ось ординат, то есть показывает значение зависимой переменной, если значение независимой переменной равно 0. Для данного примера константа не имеет большого смысла, так как масса тела не может равняться −6 327 г, да и нас не интересует масса тела новорожденных для ситуаций, когда длина стремится к нулю. Данная ситуация наглядно демонстрирует, что прогнозировать значения зависимой переменной следует только для того диапазона независимой переменной, на основании которого было построено уравнение регрессии. Так, для данного примера прогнозировать массу тела новорожденных с помощью вышеприведенного уравнения целесообразно только при значениях длины тела от 44 до 58 см, несмотря на то, что регрессионную прямую можно провести далеко за пределы имеющихся данных в оба направления. Это, кстати, является довольно частой ошибкой исследователей, когда прогнозируются значения зависимой переменной по значениям независимой переменной, которые не входили в исследование. Хотелось бы предостеречь будущих авторов от таких прогнозов, так как построенная модель не гарантирует линейной зависимости на всем протяжении регрессионной прямой.

Вышеприведенное уравнение позволяет прогнозировать только среднее значение зависимой переменной, то есть если длина тела новорожденного из генеральной совокупности, из которой была отобрана выборка, будет равна 50 см, то среднее значение массы тела всех новорожденных с длной тела 50 см будет 3 173 г. Мы сохранили (см. рис. 10) предсказанное среднее значение в виде переменной PRE_1, которое для длины 50 см. равно 3 165 г. Различия обусловлены округлением, поэтому рекомендуется использовать для дальнейших расчетов те значения, что сохраняет SPSS. Однако в каком интервале будет находиться рассчитанное среднее? Эти значения мы сохранили в файле после того, как в диалоговом окне «Linear Regression: Save» отметили Prediction intervals для средних значений (Mean). Кроме того, мы отметили Prediction intervals для индивидуальных значений массы тела (Individual), которое можно ожидать для тех значений длины тела новорожденных, которые были измерены в ходе исследования. В результате анализа SPSS создала переменные LMCI_1 и UMCI_1, показывающие нижнюю и верхнюю границы доверительного интервала для регрессионной прямой (среднего значения прогнозируемой массы тела) для каждого значения длины, включенного в данное исследование. В исследовании было несколько новорожденных с длиной тела 50 см (например, строка 13). На основании регрессионного анализа можно сказать, что для новорожденных с длиной тела 50 см среднее значение массы тела в генеральной совокупности будет с 95 % надежностью находиться в пределах от 3 146 до 3 184 г. Но можно ли говорить, что значение массы тела любого новорожденного с длиной 50 см должно попадать в этот интервал? Ответ – однозначно нет, так как интервал от 3 146 до 3 184 – это интервал, куда попадет среднее значение массы тела всех новорожденных с такой длиной. Если мы хотим спрогнозировать значение зависимой переменной и доверительные интервалы для среднего значения прогноза для тех значений независимой переменной, которые не были представлены в исследовании, например 48,5 см, то можно провести расчеты вручную, используя вышеприведенное уравнение регрессии для $\overline{\overline{Y}}$ и формулы для нижней и верхней границ доверительного интервала, представленные ниже:

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

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

$\overline{\overline{Y}}_i$ является спрогнозированным средним значением, рассчитанным по уравнению

$$\overline{\overline{Y}} = -6327 + 190X.$$

S_e в данных формулах предсталяет собой стандартное отклонение остатков, которое можно рассчитать либо вручную по формуле

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

либо с использованием SPSS путем расчета описательных статистик для остатков. Остатки были сохранены SPSS в файле в виде переменной RES_1. Каким образом рассчитывать стандартные отклонения для количественных данных, рассматривалось в [1]. В данном примере, как мы увидим ниже, стандартное отклонение остатков (S_e) равно 241 г. Под выражением $t_{(n-2)}$ подразумевается значение t для n−2 степеней свободы. Поскольку наша выборка составляет 869 человек, можно использовать значение 1,96, как для классического нормального распределения.

Однако прогноз среднего значения дает исследователю не так много. В каком интервале будут находиться значения массы тела отдельных новорожденных? Какие значения считать типичными на основании построенного уравнения регрессии? Для этого SPSS рассчитывает верхнюю и нижнюю границы предсказательного интервала, в который попадет 95 % всех значений массы тела новорожденных из генеральной совокупности. Переменные для верхней и нижней границы предсказательного интервала сохранены в виде переменных LICI_1 и UICI_1 соответственно. Так, 95 % новорожденных в генеральной совокупности с длиной тела 50 см, согласно нашим расчетам, будут иметь массу тела от 2 691 до 3 639 г. Новорожденные с длиной тела 50 см и массой менее 2 691 или более 3 639 г будут находиться за пределами 95 % предсказательного интервала для индивидуальных значений, что статистически позволит отнести их к атипичным, что не всегда может быть корректно с медицинской или биологической точки зрения. Как видно из расчетов, разброс предсказанных индивидуальных значений достаточно велик даже при относительно высоком коэффициенте детерминации модели и очень узком доверительном интервале для среднего. Последнее обусловлено достаточно большим объемом выборки. При прогнозировании индивидуальных значений зависимой переменной для тех значений независимой переменной, которые не вошли в исследование (например, 48,5 см), можно воспользоваться нижеприведенными формулами для соответственно нижней и верхней границ предсказательного интервала.

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

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

Можно запросить у SPSS построить доверительные интервалы для прогнозируемых средних и для индивидуальных наблюдений путем выбора в диалоговом окне «Properties» (см. рис. 5) в разделе «Confidence intervals» Mean или Individual, соответственно.

В следующей таблице результатов приводятся те наблюдения, остатки для которых составляют более 2 стандартных отклонений. Согласно определению нормального распределения, таких наблюдений не должно быть более 5 %. В данном примере количество наблюдений, для которых стандартизованные остатки имеют значения либо менее −2, либо более 2, составило 46 (таблица результатов, обозначенная SPSS как «Casewise diagnostics», по причине ее большого размера не приводится). От общего числа наблюдений 46 наблюдений составляют 5,3 %, что практически соответствует ожидаемым 5 %. Кроме того, количество наблюдений с большими отрицательными остатками приблизительно равно количеству наблюдений с большими положительными остатками, что говорит о симметричности распределения остатков. Наличие большого количества наблюдений с остатками более 3 стандартных отклонений говорит о большом разбросе данных вокруг регрессионной прямой, то есть о плохом соответствии модели имеющимся данным. Если большинство из «атипичных» остатков имеют один и тот же знак, то это может говорить о наличии кластера наблюдений, которые не могут быть описаны существующей моделью.

Следующая таблица результатов (Residual Statistics, рис. 15) показывает минимальные (Minimum), максимальные (Maximum) значения для различного рода остатков и вспомогательных критериев. В нашем простейшем анализе нас прежде всего интересуют значения нестандартизованных остатков (Residual), их среднее значение (Mean) и стандартное отклонение (Std. Deviation, Se), которое может пригодиться для расчета доверительных интервалов для линии регрессии и предсказательного интервала для индивидуальных значений зависимой переменной с использованием вышеприведенных формул. Для нашего примера среднее значение остатков равно 0, а стандартное отклонение – 241.

Далее результаты регрессионного анализа содержат графическую информацию. Первый график (рис. 16) представляет собой гистограмму остатков (в данном случае стандартизованных), которая должна иметь симметричную колоколообразную форму, так как одно из обязательных условий линейного

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	2025,95	5063,25	3388,20	363,122	869
Std. Predicted Value	-3,751	4,613	,000	1,000	869
Standard Error of Predicted Value	8,214	38,628	10,983	3,632	869
Adjusted Predicted Value	2021,15	5081,25	3388,23	363,267	869
Residual	-902,098	833,058	,000	240,976	869
Std. Residual	-3,741	3,455	,000	,999	869
Stud. Residual	-3,752	3,458	,000	1,001	869
Deleted Residual	-907,320	834,382	-,038	241,611	869
Stud. Deleted Residual	-3,781	3,480	,000	1,002	869
Mahal. Distance	,008	21,279	,999	1,726	869
Cook's Distance	,000	,109	,001	,005	869
Centered Leverage Value	,000	,025	,001	,002	869

a. Dependent Variable: ves

Рис. 15. Таблица анализа остатков

регрессионого анализа – нормальное распределение остатков. При невыполнении этого условия результатам регрессионного анализа верить не следует. Для данного примера распределение остатков напоминает форму нормального распределения, но с учетом того, что SPSS строит гистограммы с произвольной шириной столбцов, наибольшее доверие должно быть к квантильной диаграмме (рис. 17), которая для нашего примера представляет собой прямую линию, что говорит о нормальном распределении остатков, а значит, и о соблюдении одного из основных условий регрессионного анализа. Следующим условием для линейного регрессионного анализа является гомоскедастичность. Условие гомоскедастичности подразумевает одинаковый разброс значений зависимой переменной для любых значений независимой переменной. Мы проверяем условие гомоскедастичности путем построения графика зависимости стандартизованных остатков от стандартизованных предсказанных значений.

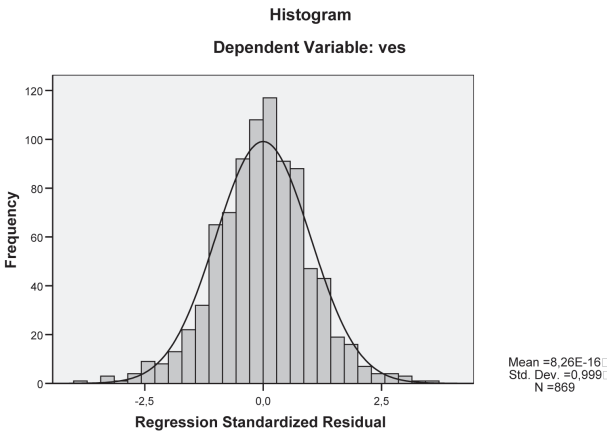

Рис. 16. Гистограмма стандартизованных остатков

О соблюдении этого условия будет говорить одинаковый разброс точек на скаттерограмме, как на рис. 18, на котором видно, что разброс остатков приблизительно одинаков для всех значений стандартизованной предсказанной величины, то есть условие гомоскедастичности соблюдается. Однако такое встречается не всегда, и график зависимости стандартизованных остатков от стандартизованных предсказанных значений может

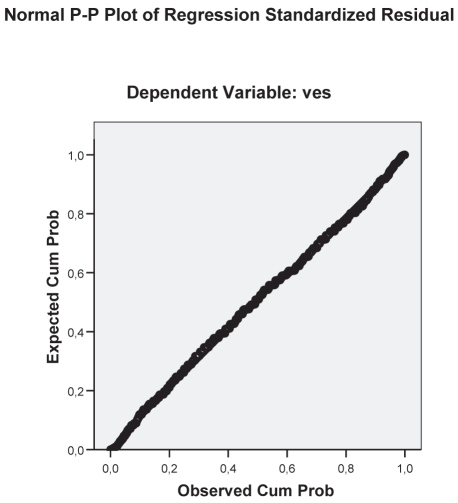

Рис. 17. Квантильная диаграмма стандартизованных остатков

иметь форму треугольника, трапеции, представлять криволинейную зависимость, и т. п. Во всех этих случаях говорят о гетероскедастичности, что не позволяет применять линейный регрессионный анализ в том виде, как мы его рассматриваем. Есть способы коррекции гетероскедастичности, с которыми читатели могут познакомиться самостоятельно в специализированной литературе.

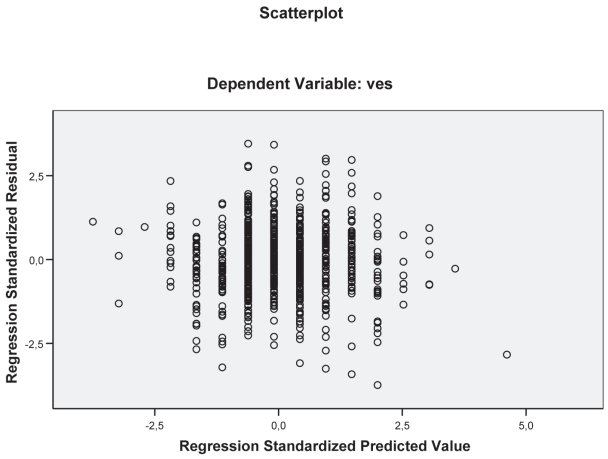

Рис. 18. Разброс стандартизованных остатков в зависимости от стандартизованных предстказанных значений

Таким образом, мы оценили коэффициент детерминации для модели, предсказывающей массу тела новорожденных по данным длины их тела, оценили статистическую значимость модели, рассчитали значения константы и коэффициентов, построили уравнение линейной регрессии, сохранили в файле средние для предсказанных значений массы тела по данным длины с 95 % доверительными интервалами для генеральных средних, сохранили значения остатков, а также 95 % предсказательный интервал для каждого из индивидуальных значений массы тела новорожденных для каждого из значений длины из нашей выборки. Прогноз как средних, так и доверительных

интервалов для средних и предсказательных интервалов для индивидуальных значений зависимой переменной для любых значений можно произвести с помощью вышеприведенных формул. Проверка соблюдения условий показала, что остатки являются независимыми, имеют нормальное распределение и одинаковый разброс на всем протяжении предсказанных значений зависимой переменной, из чего следует, что наша модель имеет достаточную внутреннюю валидность. Так как выборка является сплошной, то и внешняя валидность модели также достаточна. Внешняя валидность модели оценивается по репрезентативности выборки. Модель с высокой внутренней валидностью может быть совершенно бесполезной для генеральной совокупности, если выборка не является репрезентативной.

Перечислим еще раз необходимые условия, которые должны соблюдаться для применения линейного регрессионного анализа:

• Зависимая переменная должна быть количественной

• Независимая переменная должна быть количественной

• Наблюдения (и остатки) должны быть независимы друг от друга (проверяется с помощью критерия Durbin-Watson)

• Зависимость между переменными должна быть линейной (проверяется графически путем построения скаттерограммы)

• Остатки должны иметь нормальное распределение (проверяется графически с помощью гистограмм, квантильных диаграмм, а также с помощью критериев Kolmogorov-Smirnov или Shapiro-Wilk)

• Остатки должны иметь одинаковый разброс на всем протяжении предсказанных значений (или независимой переменной). Условие гомоскедастичности проверяется путем построения скаттерограммы, как показано на рис. 9, а результат оценивается по графику, который должен в идеальной ситуации представлять собой бессистемный разброс точек в поле зрения (рис. 18)

• Выборка должна быть репрезентативной

Еще раз хотелось бы обратить внимание читателя на то, что не следует прогнозировать значения зависимой переменной по значениям независимой переменной, которые находятся за пределами диапазона выборочных данных, так как зависимость за пределами изучаемого диапазона может сильно отличаться от линейной. Кроме того, нелишне повторить, что все условия выполнения регрессионного анализа должны тщательно проверяться перед тем, как применять результаты на практике. Например, на рис. 19–23 представлены результаты линейного регрессионного анализа для переменных X и Y. Согласно данным рис. 19–21, модель статистически значима, коэффициенты статистически значимы, то есть велик соблазн начать применять результаты моделирования на практике. Однако эта модель не является адекватной для прогнозирования, так как на рис. 21 четко видно существенное отклонение распределения остатков от нормального распределения, а на рис. 22 видно явное несоблюдение условия гомоскедастичности.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,397[a]	,158	,154	14,73826

a. Predictors: (Constant),

Рис. 19. Коэффициент детерминации для гипотетической модели (описание в тексте)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9332,199	1	9332,199	42,963	,000[a]
	Residual	49742,555	229	217,216		
	Total	59074,754	230			

a. Predictors: (Constant),

b. Dependent Variable:

Рис. 20. Таблица, показывающая статистическую значимость гипотетической модели (описание в тексте)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-36,182	7,315		-4,947	,000	-50,594	-21,770
		2,630	,401	,397	6,555	,000	1,839	3,421

a. Dependent Variable:

Рис. 21. Значения константы и коэффициента для гипотетической модели (описание в тексте)

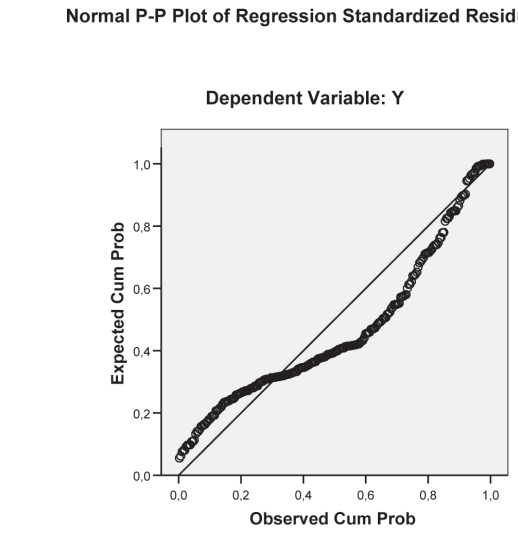

Рис. 22. Нарушение условия нормального распределения остатков для гипотетической модели (описание в тексте)

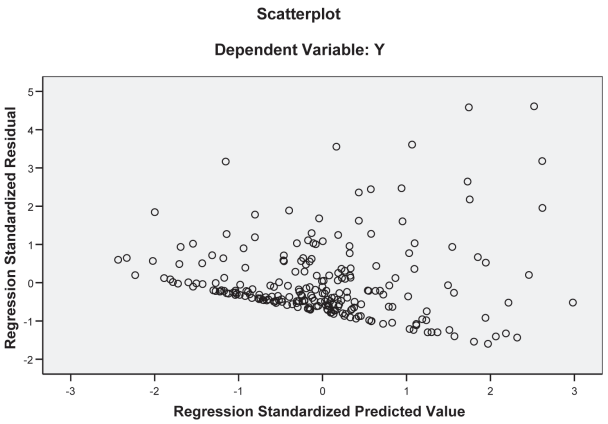

Рис. 23. Нарушение условия гомоскедастичности для гипотетической модели (объяснения с тексте)

Список литературы

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

2. *Grjibovski A. M.* Social variations in fetal growth in Northwest Russia: an analysis of medical records. / A. M. Grjibovski, L. O. Bygren, B. Svartbo, P. Magnus // Annals of Epidemiology. – 2003. – N 9. – C. 599–605.

SIMPLE LINEAR REGRESSION ANALYSIS

A. M. Grjibovski

National Institute of Public Health, Oslo, Norway

The article gives a brief description of the simple linear regression analysis and its use in situations with one dependent and one independent variable using SPSS software. The paper gives special attention to assumptions of linear regression and their testing. The article provides only general introduction to the simple method of prediction of the one variable based on the known values of the other. The readers are encouraged to consult statistical literature prior to analysing own data and preparing manuscripts.

Key words: linear regression, coefficient of determination, least squares method, confidence intervals, SPSS.

Контактная информация:

Гржибовский Андрей Мечиславович – старший советник Национального института общественного здоровья, г. Осло, Норвегия

Адрес: Nasjonalt folkehelseinstitutt, Pb 4404 Nydalen, 0403 Oslo, Norway

Тел.: +47 21076392, +47 45268913; e-mail: angr@fhi.no

Статья поступила 23.09.2008 г.