Лекция 21

ҚАЗАҚСТАННЫҢ АШЫҚ УНИВЕРСИТЕТІ

ОСНОВЫ СТАТИСТИКИ

Использование SPSS для проведения регрессионного анализа

Изучив данную лекцию, вы освоите техническую сторону использовании программы SPSS для анализа показателей уравнения регрессии как парной, так и множественной регрессии.

Простая линейная регрессия.

Чтобы вызвать регрессионный анализ в SPSS, выберите в меню Analyze... (Анализ) ► Regression... (Регрессия). Откроется соответствующее подменю.

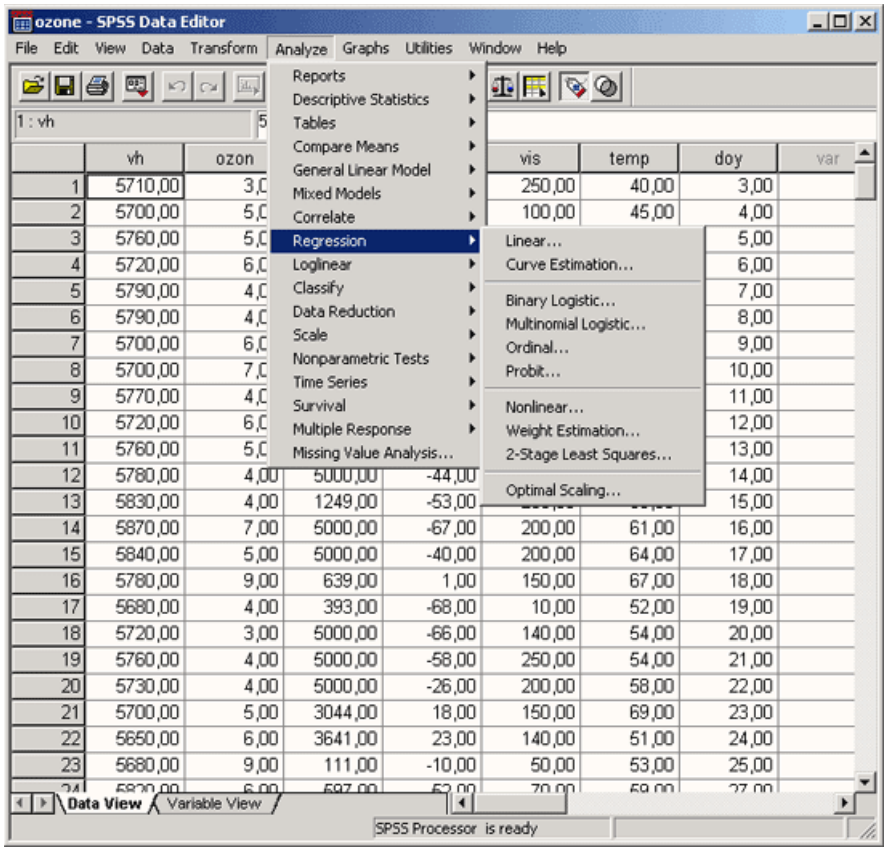

Рис. 21.1 Вспомогательное меню Regression (Регрессия)

При изучении линейного регрессионного анализа снова будут проведено различие между простым анализом (одна независимая переменная) и множественным анализом (несколько независимых переменных). Никаких принципиальных отличий между этими видами регрессии нет, однако простая линейная регрессия является простейшей и применяется чаще всех остальных видов.

При проведении простой линейной регрессии основной задачей является определение параметров b и a. Оптимальным решением этой задачи является такая прямая, для которой сумма квадратов вертикальных расстояний до отдельных точек данных является минимальной.

Расчёт уравнения регрессии.

Откройте файл, в котором размещены Ваши данные

Выберите в меню Analyze... (Анализ) ► Regression...(Регрессия) ► Linear... (Линейная). Появится диалоговое окно Linear Regression (Линейная регрессия).

Перенесите переменную chol1 в поле для зависимых переменных и присвойте переменной chol0 статус независимой переменной.

Ничего больше не меняя, начните расчёт нажатием ОК.

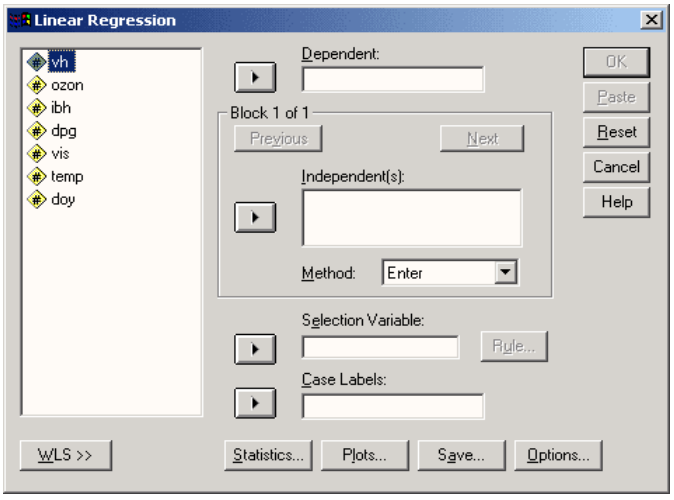

Рис.21.2 Диалоговое окно Линейная регрессия

Вывод основных результатов выглядит следующим образом: мы получим сводную таблицу по модели, которую вы сейчас видите на экране, а также таблицы дисперсионного анализа для параметров модели вы их сейчас видите на экране.

Рассмотрим сначала нижнюю часть результатов расчётов. Здесь выводятся коэффициент регрессии b и смещение по оси ординат, а под именем «константа». То есть, уравнение регрессии выглядит следующим образом:

$$chol1 = 0{,}863 \bullet chol0 + 34{,}546.$$

Частные рассчитанных коэффициентов и их стандартная ошибка дают контрольную величину T; соответственный уровень значимости относится к существованию ненулевых коэффициентов регрессии. Средняя часть расчётов отражает два источника дисперсии: дисперсию, которая описывается уравнением регрессии (сумма квадратов, обусловленная регрессией) и дисперсию, которая не учитывается при записи уравнения (остаточная сумма квадратов). Частное от суммы квадратов, обусловленных регрессией и остаточной суммы квадратов, называется «коэффициентом детерминации». В таблице результатов это частное выводится под именем «R-квадрат». В нашем примере мера определённости равна: 0,741

Эта величина характеризует качество регрессионной прямой, то есть степень соответствия между регрессионной моделью и исходными данными. Мера определённости всегда лежит в диапазоне от 0 до 1. Существование ненулевых коэффициентов регрессии проверяется посредством вычисления контрольной величины F, к которой относится соответствующий уровень значимости.

В простом линейном регрессионном анализе квадратный корень из коэффициента детерминации, обозначаемый «R», равен корреляционному коэффициенту Пирсона. При множественном анализе эта величина менее наглядна, нежели сам коэффициент детерминации. Величина «Смещенный R-квадрат» всегда меньше, чем несмещенный. При наличии большого количества независимых переменных, мера определённости корректируется в сторону уменьшения. Принципиальный вопрос о том, может ли вообще имеющаяся связь между переменными рассматриваться как линейная, проще и нагляднее всего решать, глядя на соответствующую диаграмму рассеяния.

Кроме того, в пользу гипотезы о линейной связи говорит также высокий уровень дисперсии, описываемой уравнением регрессии.

И, наконец, стандартизированные прогнозируемые значения и стандартизированные остатки можно предоставить в виде графика. Вы получите этот график, если через кнопку Plots...(Графики) зайдёте в соответствующее диалоговое окно и зададите в нём параметры все необходимые параметры в качестве переменных, отображаемых по осям y и x соответственно. В случае линейной регрессии остатки распределяются случайно по обе стороны от горизонтальной нулевой линии.

Сохранение новых переменных

Многочисленные вспомогательные значения, рассчитываемые в ходе построения уравнения регрессии, можно сохранить как переменные и использовать в дальнейших расчётах.

Для этого в диалоговом окне Linear Regression (Линейная регрессия) щёлкните на кнопке Save (Сохранить). Откроется диалоговое окно Linear Regression: Save (Линейная регрессия: Сохранение) как изображено на рисунке.

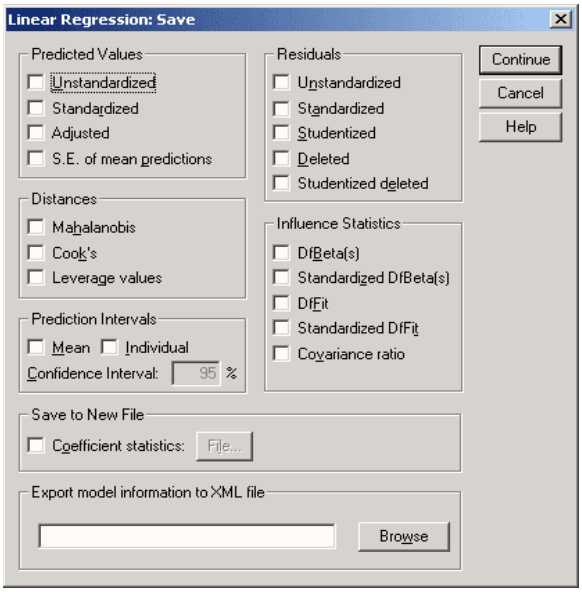

Рис. 21.3: Диалоговое окно Линейная регрессия: Сохранение.

Интересными здесь представляются опции Standardized (Стандартизированные значения) и Unstandardized (Нестандартизированные значения), которые находятся под рубрикой Predicted values (Прогнозируемые величины опции). При выборе опции Нестандартизированные значения будут рассчитывается значения y, которое соответствуют уравнению регрессии. При выборе опции Стандартизированные значения прогнозируемая величина нормализуется. SPSS автоматически присваивает новое имя каждой новообразованной переменной, независимо от того, рассчитываете ли Вы прогнозируемые значения, расстояния, прогнозируемые интервалы, остатки или какие-либо другие важные статистические характеристики. Нестандартизированным значениям SPSS присваивает имена pre_1 (predicted value), pre_2 и т.д., а стандартизированным zpr_1.

Щёлкните в диалоговом окне Linear Regression: Save (Линейная регрессия: Сохранение) в поле Predicted values (Прогнозируемые значения) на опции Unstandardized (Нестандартизированные значения).

Подтвердите нажатием Continue (Далее) и в заключение OK.

В редакторе данных будет образована новая переменная под именем pre_1 и добавлена в конец списка переменных в файле. Для объяснения значений, находящихся в переменной pre_1, возьмём случай 5. Для случая 5 переменная pre_1 содержит нестандартизированное прогнозируемое значение 263,11289. Это прогнозируемое значение слегка отличается в сторону увеличения от реального показателя содержания холестерина, взятого через один месяц (chol1) и равного 260. Нестандартизированное прогнозируемое значение для переменной chol1, так же как и другие значения переменной pre_1, было вычислено исходя из соответствующего уравнения регрессии.

Если мы в уравнение регрессии:

chol1 = 0,863 • chol0 + 34,546

подставим исходное значение для chol0 (265), то получим: chol1 = 0,863 • 265 + 34,546 = 263,241.

Небольшое отклонение от значения, хранящегося в переменной pre_1 объясняется тем, что SPSS использует в расчётах более точные значения, чем те, которые выводятся в окне просмотра результатов.

Добавьте для этого в конец файла hyper.sav, ещё два случая, используя фиктивные значения для переменной chol0. Пусть к примеру, это будут значения 282 и 314.

Оставьте предыдущие установки без изменений и проведите новый расчёт уравнения регрессии.

В конце списка переменных добавится переменная prg_2. Для нового добавленного случая (№175) для переменной chol1 будет предсказано значение 277,77567, а для случая №176 — значение 305,37620.

Построение регрессионной прямой.

Чтобы на диаграмме рассеяния изобразить регрессионную прямую, поступите следующим образом:

Выберите в меню следующие опции Graphs ... (Графики) / Scatter plots... / Диаграммы рассеяния. Откроется диалоговое окно Scatter plots... (Диаграмма рассеяния) –

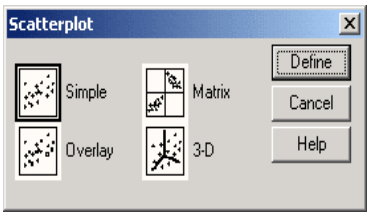

Рис. 21.4 Диалоговое окно Scatter plots... (Диаграмма рассеяния).

В диалоговом окне Scatter plots...(Диаграмма рассеяния) оставьте предварительную установку Simple (Простая) и щёлкните на кнопке Define (Определить). Откроется диалоговое окно Simple Scatter plot (Простая диаграмма рассеяния) (рис. 16.5).

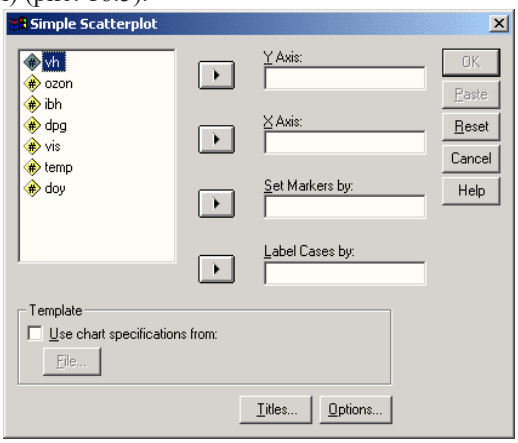

Рис. 21.5 Диалоговое окно Simple Scatterplot (Простая диаграмма рассеяния).

Перенесите переменную chol1 в поле оси Y, а переменную chol0 в поле оси X.

Подтвердите щелчком на OK. В окне просмотра результатов появится диаграмма рассеяния (рис. 21.6).

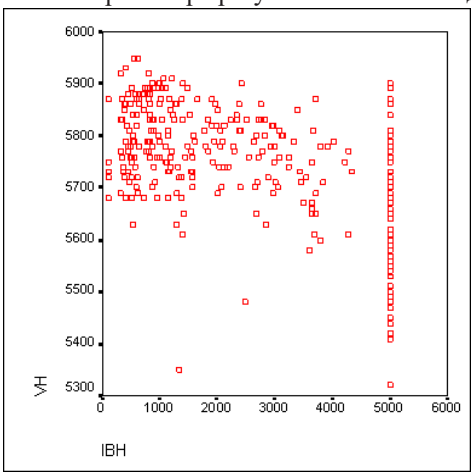

Рис. 21.6 Диаграмма рассеяния в окне просмотра

Щёлкните дважды на этом графике, чтобы перенести его в редактор диаграмм.

Выберите в редакторе диаграмм меню Chart... (Диаграмма) / Options... (Опции). Откроется диалоговое окно Scatterplot Options (Опции для диаграммы рассеяния) (рис. 21.7).

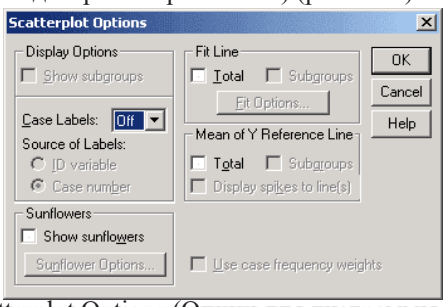

Рис. 21.7 Диалоговое окно Scatterplot Options (Опции для диаграммы рассеяния).

В рубрике Fit Line (Приближенная кривая) поставьте флажок напротив опции Total (Целиком для всего файла данных) и щёлкните на кнопке Fit Options (Опции для приближения). Откроется диалоговое окно Scatterplot Options: Fit Line (Опции для диаграммы рассеяния: приближенная кривая) (см. рис. 21.8).

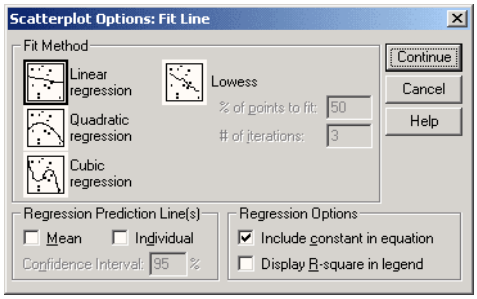

Рис. 21.8 Диалоговое окно Scatterplot Options: Fit Line (Опции для диаграммы рассеяния:

Подтвердите предварительную установку Linear Regression (Линейная регрессия) щелчком Continue (Далее) и затем на OK.

Закройте редактор диаграмм и щёлкните один раз где-нибудь вне графика. Теперь в диаграмме рассеяния отображается регрессионная прямая (рис. 21.9).

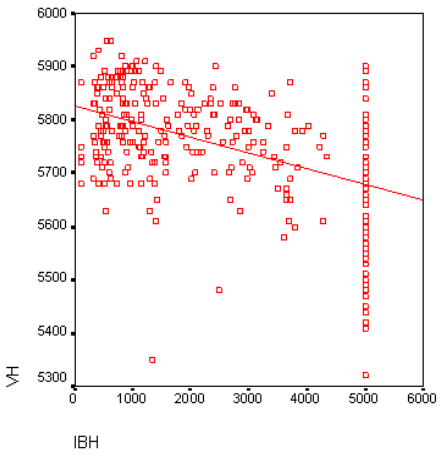

Рис. 21.9 Диаграмма рассеяния с регрессионной прямой

Выбор осей.

Для диаграмм рассеяния часто оказывается необходимой дополнительная корректировка осей. Продемонстрируем такую коррекцию при помощи одного примера. В файле raucher.sav находятся десять фиктивных наборов данных. Переменная konsum указывает на количество сигарет, которые выкуривает один человек в день, а переменная pulsha количество времени, необходимое каждому испытуемому для восстановления пульса до нормальной частоты после двадцати приседаний. Как было показано ранее, постройте диаграмму рассеяния с внедрённой регрессионной прямой.

В диалоговом окне Simple Scatterplot (Простая диаграмма рассеяния) перенесите переменную puls в поле оси Y, а переменную konsum — в поле оси X. После соответствующей обработки данных в окне просмотра появится диаграмма рассеяния, изображённая на рисунке

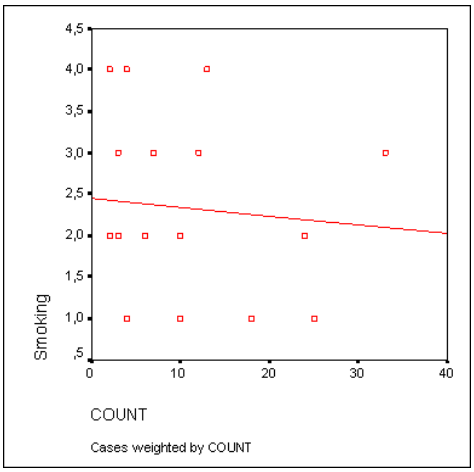

Рис. 21.10 Диаграмма рассеяния с регрессионной прямой до коррекции осей.

Дважды щёлкните на графике и в меню редактора диаграмм вберите опции Chart... (Диаграмма) / Axis... (Оси). Откроется диалоговое окно Axis Selection (Выбор оси) (рис. 21.11).

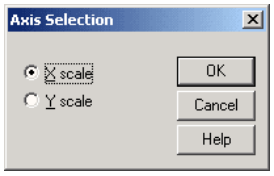

Рис. 21.11 Диалоговое окно Axis Selection (Выбор оси).

Подтвердите предварительный выбор оси X нажатием кнопки OK. Откроется диалоговое окно X-Scale Axis (Ось X) (рис. 21.12).

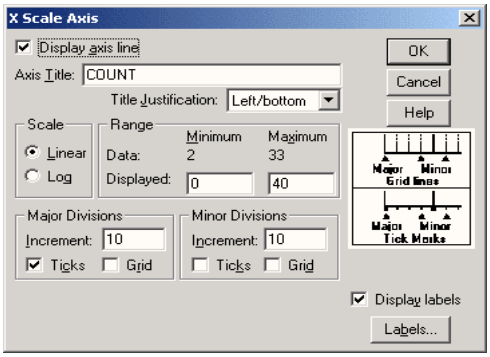

Рис. 21.12 Диалоговое окно X-Scale Axis (Ось X).

В редактируемом поле Displayed (Отображаемый) в рубрике Range (Диапазон) измените минимальное значение на 0.

Подтвердите нажатием на OK.

Выберите вновь в меню редактора диаграмм опции Chart... (Диаграмма* Axis... (Оси).

Активируйте в диалоговом окне Axis Selection (Выбор оси) опцию Y Scale (Ось Y). Откроется диалоговое окно Y-Scale Axis (Ось Y).

И здесь в рубрике Range (Диапазон) в редактируемом поле Displayed (Отображаемый) измените минимальное значение на «0».

Подтвердите нажатием на OK.

В окне просмотра Вы увидите откорректированную диаграмму рассеяния (см. рис. 21.13).

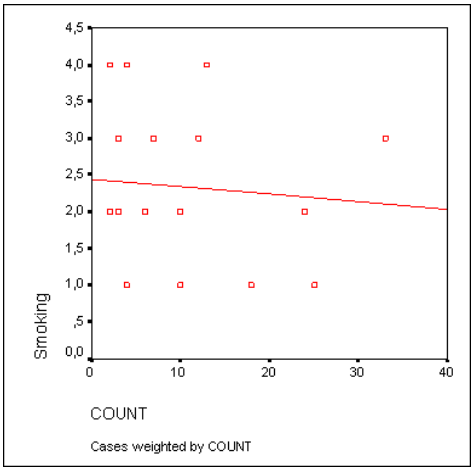

Рис. 21.13 Диаграмма рассеяния с регрессионной прямой после корректировки осей.

На откорректированной диаграмме рассеяния теперь стало проще распознать начальную точку на оси Y, которая образуется при пересечении с регрессионной прямой. Значение этой точки примерно равно 2,9. Сравним это значение с уравнением регрессии для переменных puls (зависимая переменная) и konsum (независимая переменная). В результате расчёта уравнения регрессии в окне отображения результатов появятся следующие значения:

a. Dependent Variable: Pulsfrequenz unter 80 (Зависимая переменная: равная 80) Что дает следующее уравнение регрессии:

puls = 0,145 • konsum + 2,871.

Константа в вышеприведенном уравнении регрессии (2,871) соответствует точке на оси Y, которая образуется в точке пересечения с регрессионной прямой.

В общем случае в регрессионный анализ вовлекаются несколько независимых переменных. Это, конечно же, наносит ущерб наглядности получаемых результатов, так как подобные множественные связи в конце концов становится невозможно представить графически.

В случае множественного регрессионного анализа речь идёт необходимо оценить коэффициенты уравнения

y = b1 • x1 + b2 • x2+ ... + bn • xn + a,

где n — количество независимых переменных, обозначенных как x1 и xn, a — некоторая константа.

Переменные, объявленные независимыми, могут сами коррелировать между собой; этот факт необходимо обязательно учитывать при определении коэффициентов уравнения регрессии для того, чтобы избежать ложных корреляций.

В качестве примера рассмотрим стоматологическое обследование 1130 человек, в котором исследуется вопрос необходимости лечения зубного ряда, измеряемой при помощи так называемого показателя CPITN, в зависимости от набора различных переменных. Получим таблицу, которая содержит, расшифровку переменных.

Переменные cpitn и alter принадлежат к интервальной шкале, а переменные s, pu и zb при более подробном рассмотрении можно отнести к порядковой (ранговой) шкале, так что они могут быть подвергнуты регрессионному анализу. Переменная g относится к номинальной шкале, но в то же время является дихотомической. Поэтому если при оценке результатов обратить внимание на полярность, то и

эта переменная так же может быть вовлечена в регрессионный анализ. Однако, переменная beruf относится к номинальной шкале и имеет более двух (а именно четыре) категории. Поэтому, без дополнительной обработки ее нельзя применять в дальнейших расчётах.

Для множественного анализа с несколькими независимыми переменными не рекомендуется оставлять метод включения всех переменных (Enter), установленный по умолчанию. Этот метод соответствует одновременной обработке всех независимых переменных, выбранных для анализа, и поэтому он может рекомендоваться для использования только в случае простого анализа с одной независимой переменной. Для множественного анализа следует выбрать один из пошаговых методов.

В списке Method имеются следующие возможности:

1. Enter – простейший способ - все данные формируются в единую группу.
2. Remove – это метод, который позволяет отбрасывать переменные в процессе определения конечной модели.
3. Stepwise – это метод, который позволяет добавлять и удалять отдельные переменные в соответствии с параметрами, установленными в окне Options.
4. Backward – данный метод позволяет последовательно удалять переменные из модели в соответствии с параметрами в окне Options, до того момента, пока это возможно (например по критерию значимости).
5. Forward – данный метод позволяет последовательно добавлять переменные в модель в соответствии с параметрами в окне Options, до того момента, пока это возможно.

При прямом методе независимые переменные, которые имеют наибольшие коэффициенты частичной корреляции с зависимой переменной пошагово увязываются в регрессионное уравнение. При обратном методе начинают с результата, содержащего все независимые переменные и затем исключают независимые переменные с наименьшими частичными корреляционными коэффициентами, пока соответствующий регрессионный коэффициент не оказывается незначимым (в данном случае уровень значимости равен 0,1).

Наиболее распространенным является пошаговый метод, который устроен так же, как и прямой метод, однако после каждого шага переменные, используемые в данный момент, исследуются по обратному методу. При пошаговом методе могут задаваться блоки независимых переменных; в этом случае заданные блоки на одном шаге обрабатываются совместно.

Выберите пошаговый метод, но воздержитесь от блочной формы ввода данных, не задавайте больше ни каких дополнительных расчётов и начните вычисление нажатием ОК, тогда сводная модель таблицы будет иметь вид

Model (Модель)	R	R Square (Коэффициент детерминации)	Adjusted R Square (Скорректированный R-квадрат)	Std. Error of the Estimate (Стандартная ошибка оценки)
1	,452[a]	,204	,203	,8316
2	,564[b]	,318	,317	,7698
3	,599[c]	,359	,358	,7467
4	,609[d]	,371	,369	,7402
5	,613[e]	,375	,373	,7380

Из первой таблице следует, что вовлечение переменных в расчет производилось за пять шагов. Для каждого шага происходит вывод коэффициентов множественной регрессии, меры определённости, смещенной меры определённости и стандартной ошибки.

Вдобавок ко всему для каждого шага анализируются исключённые переменные. В приведенной таблице в объяснениях нуждаются лишь коэффициенты ß. Это – регрессионные коэффициенты, стандартизованные соответствующей области значений, они указывают на важность независимых переменных, вовлечённых в регрессионное уравнение.

Уравнение регрессии для прогнозирования значения выглядит следующим образом:
cpitn = 0,032 • alter - 0.379 • pu + 0,229 • zb - 0,083 • s + 0,143 • benuf2 + 2,022.

Для 40-летнего рабочего с неполным школьным образованием, который ежедневно чистит зубы один раз в день и меняет щётку раз в полгода, с учётом соответствующих кодировок, получается следующее уравнение:

cpitn = 0,032 • 40 - 0,379 • 2 + 0,229 • 3 - 0,083 • 2 + 0,143 • 1 + 2,022 = 3,208

При помощи соответствующих опций можно организовать вывод большого числа дополнительных статистических характеристик и графиков, на которых мы здесь останавливаться не будем. Можно также создать много дополнительных переменных и добавить их в исходный файл данных.

Коллинеарность.

Важным шагом перед запуском процедуры построения регрессионной модели может быть пункт Collinearity Diagnostics в диалоговом окне Statistics.... Установление требования провести диагностику наличия коллинеарности между независимыми переменными позволяет избежать эффекта мультиколлинеарности, при котором несколько независимых переменных могут иметь настолько сильную корреляцию, что в регрессионной модели обозначают, в принципе, одно и то же (это неприемлемо).

Результат диагностики коллинеарности показан в таблице Coefficients в колонках Collinearity Statistics. Если величина значения VIF (Variance Inflation Factor) возле каждой независимой переменной меньше 10 — значит, эффекта мультиколлинеарности не наблюдается и регрессионная модель приемлема для дальнейшей интерпретации. Чем выше показатель VIF, тем более связаны между собой переменные. Если какая-либо переменная превышает значение в 10 VIF, следует пересчитать регрессию без этой независимой переменной.

Анализ остатков.

Важным моментом является анализ остатков, то есть отклонений наблюдаемых значений от теоретически ожидаемых. Остатки должны появляться случайно (то есть не систематически) и подчиняться нормальному распределению. Это можно проверить, если с помощью кнопки Charts... (Диаграммы) построить гистограмму остатков. В приведенном примере наблюдается довольно хорошее согласование гистограммы остатков с нормальным распределением.

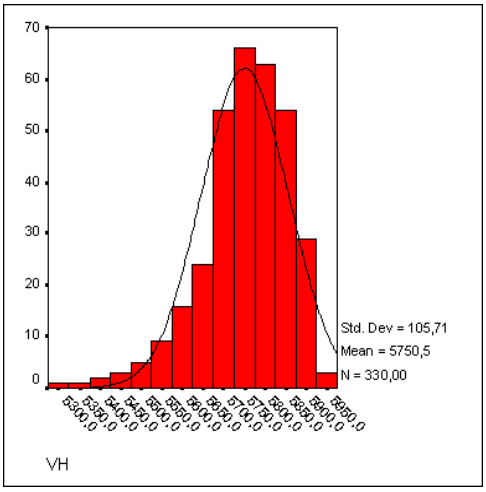

Рис. 21.14 Гистограмма остатков