**ЛЕКЦИЯ 4 РЕГРЕССИОННЫЙ АНАЛИЗ**

*1 Понятие о регрессии*

*2 Ошибки элементов уравнения прямолинейной регрессии*

*3 Реализация прямолинейной регрессии в пограммных пакетах Excel и STATISTICA.*

**1 Понятие о регрессии**

Прямолинейная корреляция характеризуется тем, что при этой форме связи каждому из одинаковых изменений первого признака соответствует вполне определенное и также одинаковое в среднем изменение другого признака, связанного с первым или зависящего от первого.

Величина, на которую в среднем изменяется второй признак, при изменении первого на единицу измерения – это **коэффициент регрессии**. Рассчитывается он по формуле.



где **R2/1** – коэффициент регрессии второго признака по первому; **σ2** – стандартное отклонение второго признака, который изменяется в связи с изменением первого; **σ1** – стандартное отклонение первого признака, в связи с изменением которого изменяется второй признак; ***r1-2*** – коэффициент корреляции между первым и вторым признаками.

*Ошибка коэффициента* регрессии равна ошибке коэффициента корреляции, умноженной на отношение сигм.



*Критерий достоверности* коэффициента регрессии равен критерию достоверности коэффициента корреляции:



**Пример**

Для разработки способа определения веса лошадей без взвешивания по обхвату груди было взвешено 1618 лошадей и у каждой из них измерен обхват груди. Получены следующие показатели: *х* – обхват груди:

*n*= 1618, *μ****х*** = 174 см, σ***x***= 7,9 см;

*n =* 1618*, μy =* 424 кг, σ*y =* 56,8 кг.

Коэффициент корреляции rx/y = +0,89±0,011;

Коэффициент регрессии массы по обхвату равен:



Ошибка коэффициента регрессии массы лошадей по обхвату их груди равна:



Достоверность этого коэффициента регрессии определяется следующим образом:

, * =* 1618–2 =1616*, tst =* {2,0 – 2,6 – 3,3}

Возможная максимальная погрешность при прогнозе генерального параметра:

* = tst х sR =* 2,0 х 0,08 = 0,16

Доверительные границы:



Таким образом, можно ожидать, что при увеличении (или уменьшении) обхвата груди на 1 см вес лошадей увеличится (или уменьшится) в среднем на R = 6,4 кг при гарантированном минимуме изменения +6,24 кг и возможном максимуме +6,56 кг, если учитывать изменения признаков в обе стороны от их средней величины.

Коэффициент прямолинейной регрессии показывает, на сколько от своей средней отклоняется второй признак, если первый признак от своей средней отклонился на единицу измерения. Это можно выразить следующей формулой:

(*X2 – μ2) = R2/1 (X1 – μ1)*

Обозначая X1 через ***х***, X2 через ***у***, R1/2 через b и произведя необходимые преобразования этого выражения, можно получить рабочую формулу прямолинейной регрессии:

*y*=*a*+*bx,*

.

По этой формуле, зная значение ***х*** (аргумент), можно определить значение ***у*** (функция) без непосредственного его измерения: нужно аргумент ***х*** помножить на коэффициент регрессии и к полученному произведению прибавить (или отнять) свободный член ***а***.

Для предыдущего примера (определение веса лошадей по обхвату груди) уравнение регрессии может быть выведено следующим образом:

*а = μу–Ry/x⋅μx* = 424–(+6,4) 174 = –690,

*b = Ry/x*= +6,4,

*у = а + bх* = –690 + 6,4*x.*

Следовательно, чтобы определить (без взвешивания живой вес лошади по этому способу, надо обхват груди лошади умножить на постоянный коэффициент 6,4 и из полученного произведения вычесть постоянное число – 690.

На основе уравнения прямолинейной регрессии можно заранее рассчитать значение функции для каждого значения аргумента. По обхвату груди можно определить живой вес лошадей.

Если эти цифры нанести на график, по оси абсцисс которого отложить через равные интервалы значения аргумента (обхвата), а по оси ординат – значения функции (веса), то получится номограмма для определения веса лошадей без взвешивания и без вычислений.

**2 Ошибки элементов уравнения прямолинейной регрессии**

В уравнении простой прямолинейной регрессии *у=а+bх* возникают 3 ошибки репрезентативности:

1) Ошибка коэффициента регрессии:



2) Ошибка уравнения регрессии, т. е. ошибка средней величины функции для каждого значения аргумента:



По данным вышеприведенного примера:



Следовательно, максимальная погрешность в определении уровня точек линии регрессии при первом пороге вероятности безошибочных прогнозов (β1 = 0,95, t1=2,0) будет равна:

*= t⋅* *=* 2⋅0,62кг *=* ± 1,24 кг

3) Ошибка индивидуальных определений функции:



Для вышеприведенного примера:

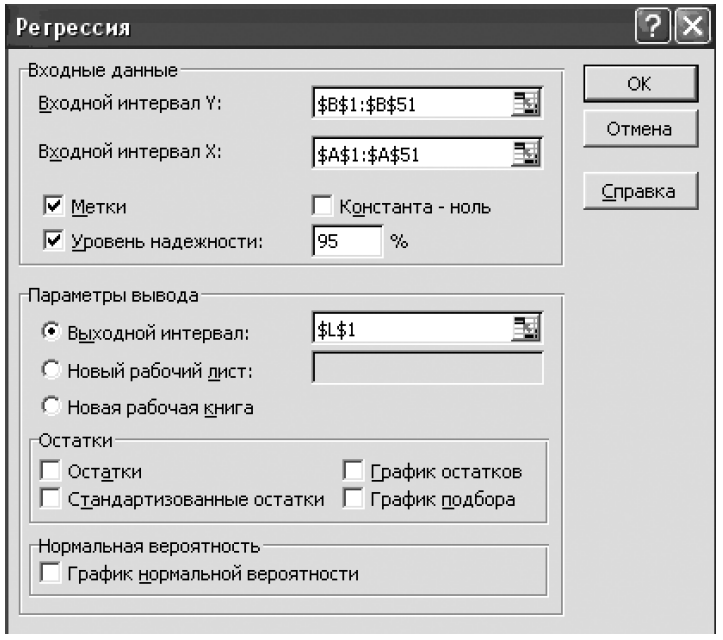


Следовательно, индивидуальная погрешность в определении веса лошадей по обхвату груди по найденной формуле регрессии, принимая первый порог вероятности безошибочных прогнозов (β1=0,95, t1=2,0), в крайних случаях не будет превышать Δ=2х26 кг=±52 кг.

**3 Реализация прямолинейной регрессии в пограммных пакетах Excel и STATISTICA**

*3.1 Регрессионный анализ в Excel*

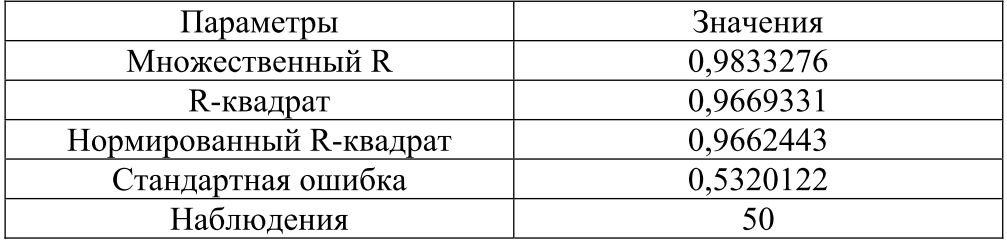
Для выполнения регрессионного анализа необходимо ввести данные также, как и для корреляционного анализа. Затем открыть модуль **Анализ данных** и выбрать опцию **Регрессия**, после чего щелкнуть мышкой **OK**. В появившемся окне выполнить установки, как показано на рис. 1, определив необходимые интервалы. Щелкнуть мышкой **OK**.



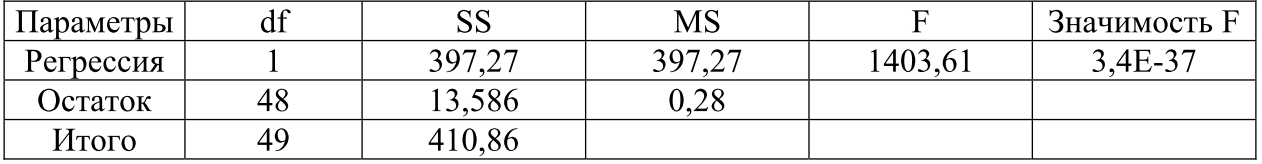
*Рис. 1*

Результат обработки появится в указанном вами месте (рабочем листе, новом листе, новой книге):

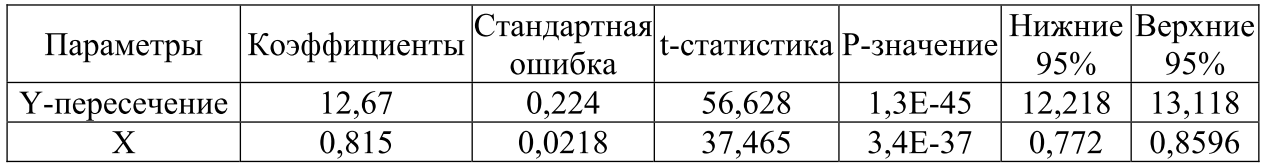
***Результат обработки***



***Дисперсионный анализ***



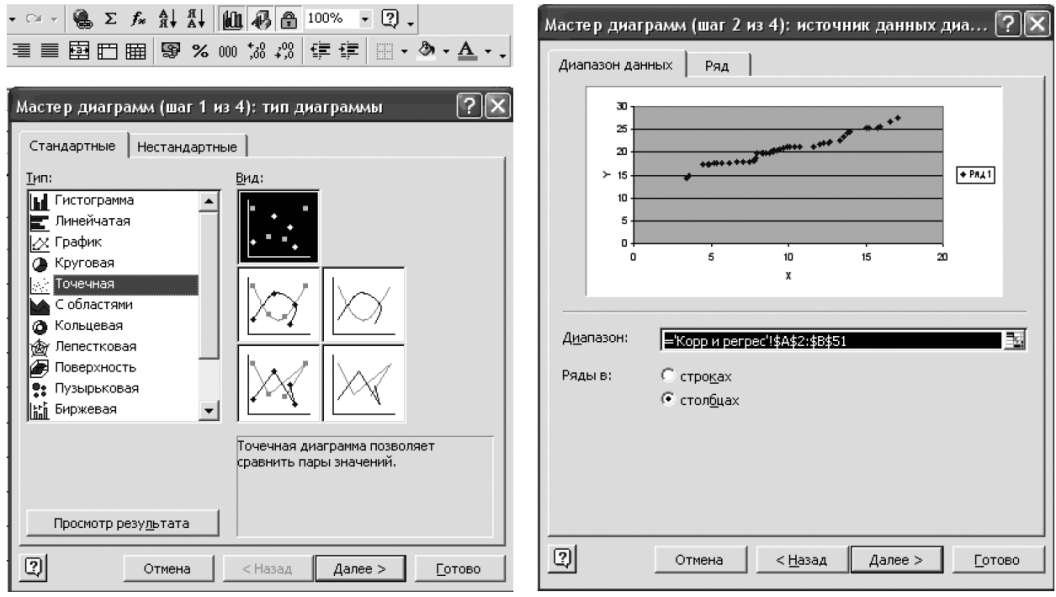
***Регрессионный анализ***



Для данного массива данных получена очень надежная регрессия с высоким коэффициентом корреляции:

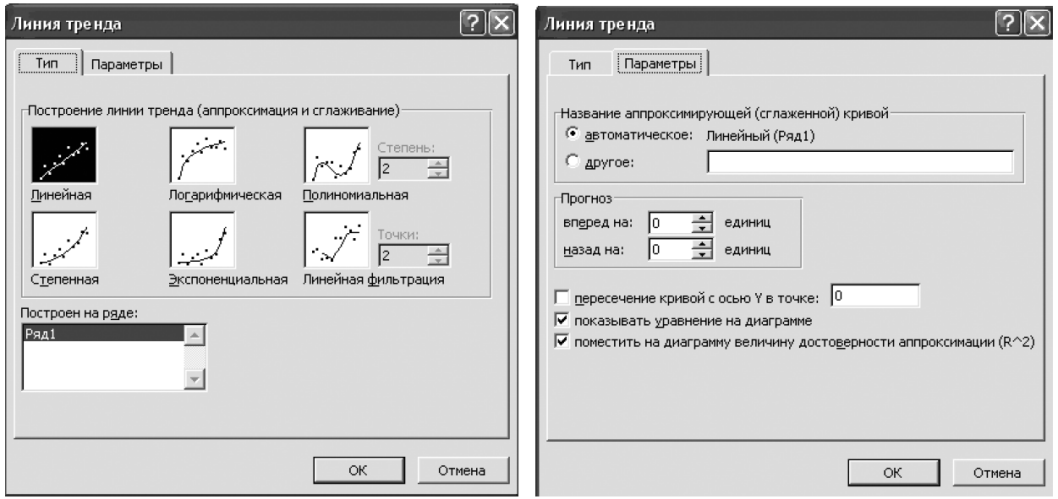
**Y=12,668+0,8158\*X; r=0,9833276**

Операции регрессии и корреляции можно выполнить в системе EXCEL, используя модуль **Мастер диаграмм**. В системе EXCEL открыть модуль **Мастер диаграмм** (рис. 2). Выбрать **Тип диаграммы** **Точечная**. Щелкнуть по кнопке **Далее**. Выбрать диапазон данных (рис. 3), оформить график и нажать **Готово**:



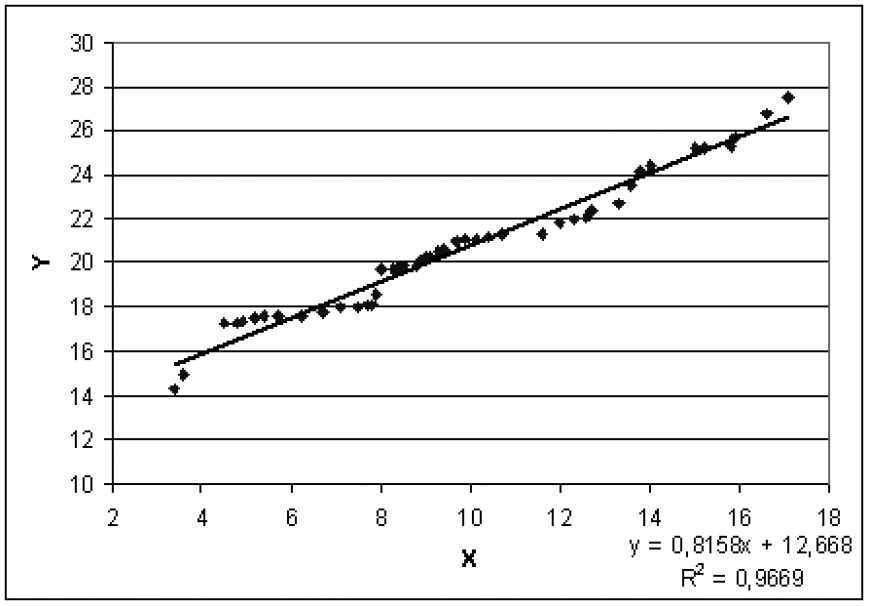
*Рис. 2 – Стартовая панель Рис. 3 – Диапазон данных*

Затем, уже в готовой диаграмме щелкнуть правой кнопкой по точкам и выбрать опцию **Добавить** **линию** **тренда**. Выбрать **Тип** – **линейная** (рис. 4). В опции **Параметры** выбрать установки, как показано на рис. 5.



*Рис. 4 – Тип аппроксимации Рис. 5 – Параметры кривой*

Отредактированная диаграмма представлена на рис. 6:



*Рис. 6 – Отредактированная диаграмма*

Уравнение регрессии и R2 находятся в правом нижнем углу диаграммы. Как видно, что они такие же, как и при выполнении регрессионного анализа.

*3.1 Регрессионный анализ в STATISTICA*

Регрессионный анализ в системе STATISTICA проводится в модуле **Multiple regression (множественная регрессия)**.

***Описание модели***

В рамках линейной регрессионной модели описывается зависимость Y от Х. При этом постулируется, что наблюдаемые величины связаны между собой регрессионной зависимостью вида:

**Y(i) = B1\* X(i) + В0 +e(i), при 0 < I < = n**,

где **B1**, **В0** – неизвестные константы, **e(i)** – ненаблюдаемые случайные величины (наблюдаются только **X(i)**, **Y(i)**, **0<i<=n**) со средним **0** (несмещенные) и неизвестной дисперсией, не меняющейся от опыта к опыту.

Иногда случайные величины **e(i), 0<i<=n** называют ошибками наблюдения. Относительно **e(i)** предполагается, что они не коррелированы в разных опытах. Кроме того, часто предполагается, что ошибки имеют нормальное распределение. В этом случае некоррелированность влечет независимость.

Общая задача регрессионного анализа состоит в том, чтобы по наблюдениям **(X(1),Y(2)), ... (X(n),Y(n))**:

- оценить параметры модели **В1, В0** наилучшим образом;

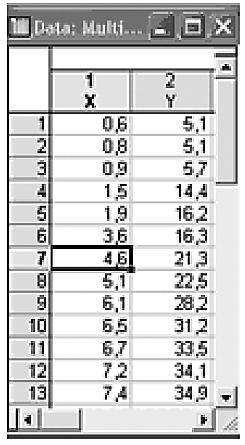
- построить доверительные интервалы для **В1, В0**;

- проверить гипотезу о значимости регрессии;

- оценить степень адекватности модели и т. д.

Рассмотрим на конкретном примере, как решается данная задача в системе STATISTICA.

Для начала необходимо ввести исходные данные (рис. 7):

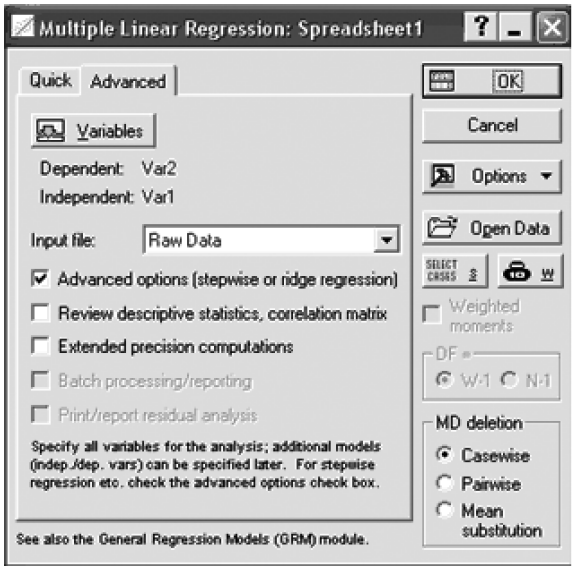


*Рис. 7 – Исходные данные*

Здесь: **X** – независимая переменная; **Y** – зависимая переменная

**Шаг 1.** Из **Переключателя модулей** STATISTICA откройте модуль **Multiple regression (множественная регрессия)**: высветите название модуля и далее щелкните мышью по названию модуля **Multiple regression**.

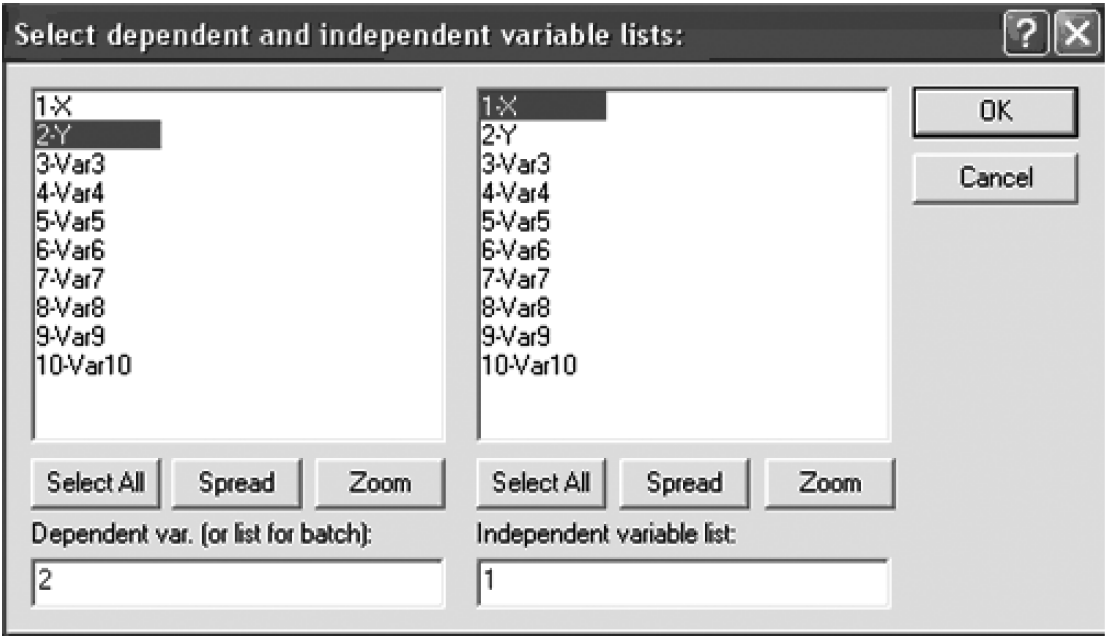
**Шаг 2.** На экране появится стартовая панель модуля (рис. 8):



*Рис. 8 – Стартовая панель модуля Multiple regression*

Выполните установки, как показано на рис. 8. Далее выберите переменные для анализа с помощью кнопки **Variables (переменные)**.

После того как кнопка будет нажата, на вашем экране появится диалоговое окно **Select dependent and independent variable list (выбрать списки зависимых и независимых переменных)** (рис. 9):

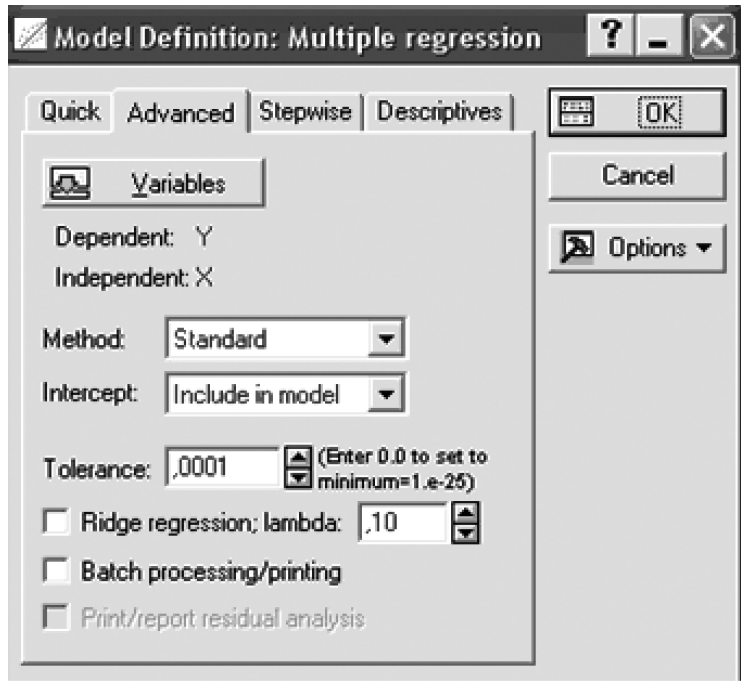


*Рис. 9 – Окно выбора переменных для анализа*

Высветив имя переменной в левой части окна, выберите зависимую переменную. Высветив имя переменной в правой части окна, выберите независимую переменную (напомним, что в данном примере независимой переменной является X, зависимой – Y).

Высветив эти переменные, как показано на рисунке, нажмите кнопку **ОК**. Вы вновь окажетесь в стартовой панели модуля. Нажмите кнопку **ОК** в правом углу стартовой панели.

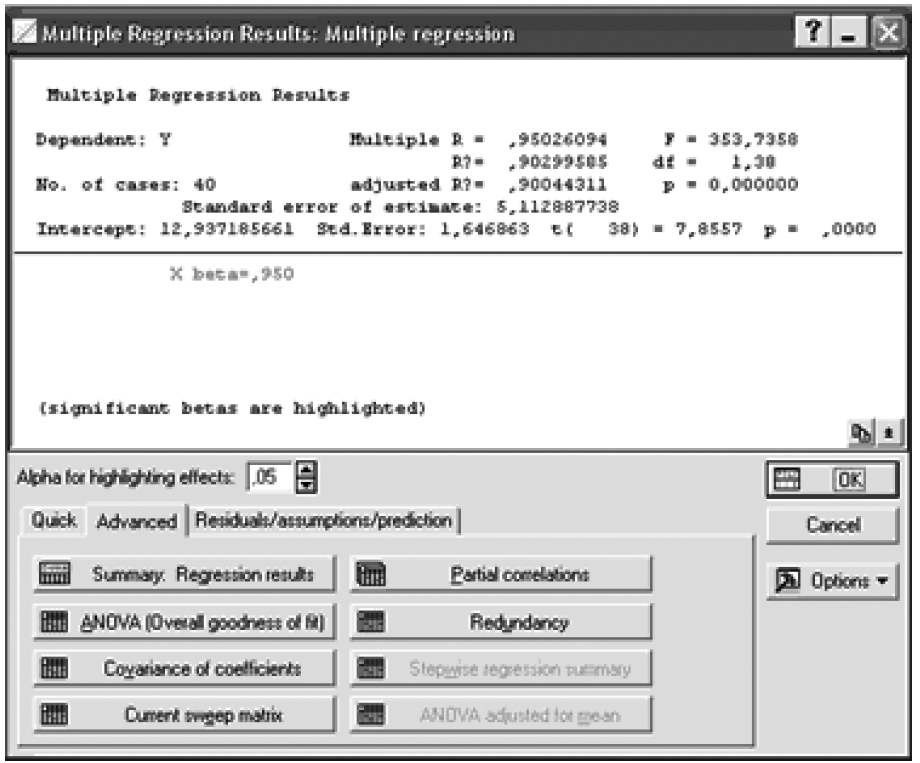
**Шаг 3**. На экране появится диалоговое окно **Model Definition (построение модели)** (рис. 10):



*Рис. 10 – Окно построения модели в модуле* ***Множественная регрессия***

В данном окне выберите стандартный метод оценивания, в опции **Method (метод)**: **Standard (стандартный)**. Далее нажмите кнопку **OK**. Программа произведет оценивание параметров модели и на экране появится диалоговое окно результатов.

**Шаг 4**. В диалоговом окне **Multiple Regression Results (результаты множественной регрессии)** просмотрите результаты оценивания (рис. 11). Результаты можно просмотреть в численном и графическом виде. Окно результатов анализа имеет следующую структуру: верх окна – информационный. Он состоит из двух частей: в первой части содержится основная информация о результатах оценивания, во второй высвечиваются значимые регрессионные коэффициенты. Внизу окна **Результаты множественной регрессии** находятся функциональные кнопки, позволяющие всесторонне просмотреть результаты анализа.



*Рис. 11 – Окно оценивания параметров*

Рассмотрим вначале информационную часть окна. В ней содержатся краткие сведения о результатах анализа:

- **Dep. Var.** (имя зависимой переменной). В нашем случае – Y.

- **No. of Cases** (число случаев, по которым построена регрессия). В нашем примере это – 40.

- **Multiple R** (коэффициент множественной корреляции).

- **R2** (квадрат коэффициента множественной корреляции), обычно называемый коэффициентом детерминации. Это чрезвычайно важная характеристика, одна из основных статистик в данном окне и показывает долю общего разброса (относительно выборочного среднего зависимой переменной), которая объясняется построенной регрессией.

- **Adjusted R2** (скорректированный коэффициент детерминации), определяемый как:

**Adjusted R2 = R2 \* n/(n-p)**,

где n – число наблюдений в модели, р – число параметров модели (число независимых переменных плюс 1, так как в модель включен свободный член).

- **Std. Error of estimate (стандартная ошибка оценки)**. Это статистика является мерой рассеяния наблюдаемых значений относительно регрессионной прямой.

- **Intercept (оценка свободного члена регрессии)**. Значение коэффициента **В0** в уравнении регрессии.

- **Std. Error** (стандартная ошибка оценки свободного члена). Стандартная ошибка коэффициента **В0** в уравнении регрессии.

- **t(df) and p-value** (значение t-критерия и уровень р). t-критерий используется для проверки гипотезы о равенстве 0 свободного члена регрессии.

- **F** – значения F-критерия;

− **df** – число степеней свободы F-критерия;

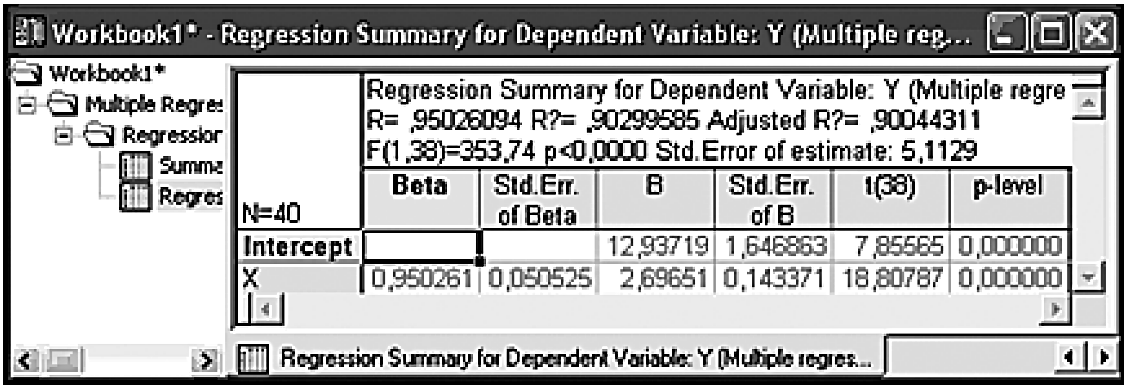
− **р** – уровень значимости.

В информационной части обратим внимание, прежде всего, на значения коэффициента детерминации. Значения коэффициента детерминации лежат в пределах от 0 до 1. В нашем примере R2 = 0,90... Это очень хорошее значение, показывающее, что построенная регрессия объясняет более 90% разброса значений переменной X относительно среднего.

Далее посмотрите на значение F-критерия и уровень значимости р. F-критерий используется для проверки гипотезы о значимости регрессии. В данном случае для проверки гипотезы, утверждающей, что между зависимой переменной X и независимой переменной Y нет линейной зависимости, то есть **В1=0**, против альтернативы **В1** не равен **0**. В данном примере наблюдается большое значение F-критерия –353,7358 и даваемый в окне уровень значимости р = 0,0000, показывающие, что построенная регрессия высоко значима.

Рассмотрим вторую часть информационного окна. В этой части система сама говорит нам о значимых регрессионных коэффициентах, высвечивая строку: **X beta = 0,950** и на пояснение **significant beta's are highlighted (значимые beta высвечены)**. Отметим, что в данном случае **beta** – стандартизованный коэффициент **В1** (коэффициент при независимой переменной X).

Перейдем в функциональную часть окна результатов. Прежде всего, нажмите кнопку **Regression summary (итоговый результат регрессии)**. На экране появится электронная таблица вывода, в которой представлены итоговые результаты оценивания регрессионной модели (рис. 12).



*Рис. 12 – Итоговая таблица регрессии*

В первом столбце таблицы даны значения коэффициентов **beta – стандартизованные коэффициенты регрессионного уравнения**, во втором – стандартные ошибки **beta**, в третьем – точечные оценки параметров модели.

Свободный член **В0 = 12,93719**.

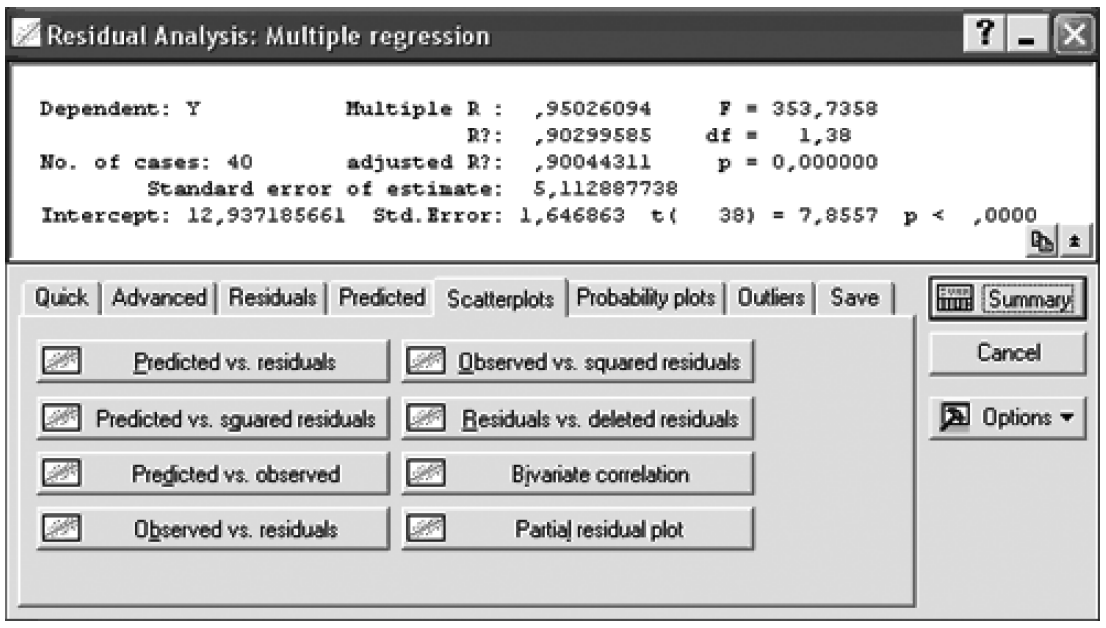
Коэффициент **В1** (при независимой переменной **X**) **= 2,69651**.

Далее, стандартные ошибки для **В0, В1**, значения **t-критерия** и т. д.

Из таблицы видно, что оцененная модель имеет вид:

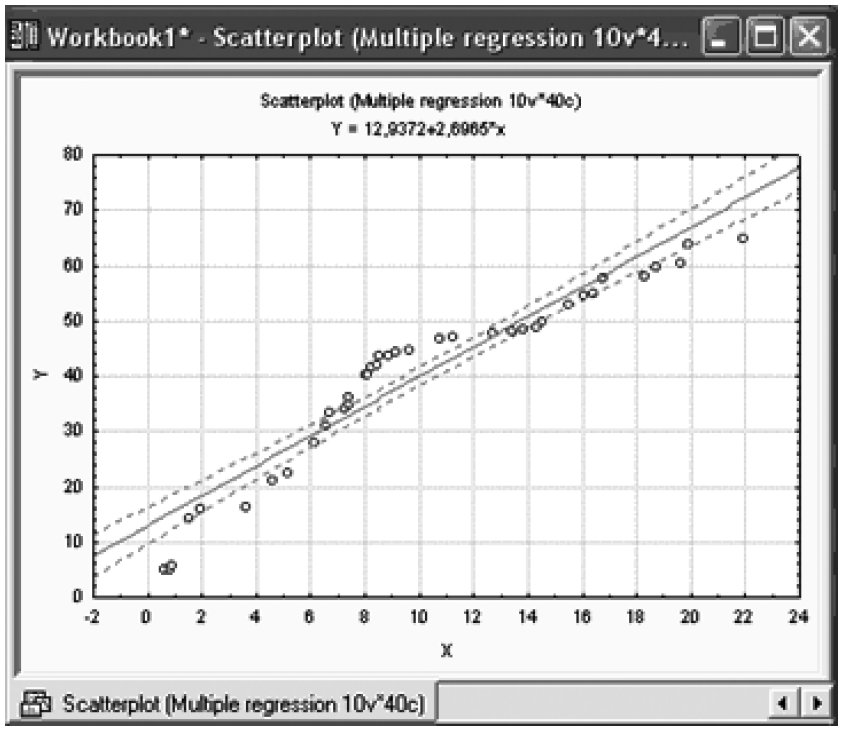
**Y = 2,69651\*X +12,93719**

**Шаг 5**. В окне оценивания параметров (рис. 11) последовательно нажать кнопки **OK** (рис. 13), **Scatterplot**, **Bivariate Correlation**. В появившемся окне выбрать зависимую (**Y variable**) и независимую (**X variable**) переменные. Нажать кнопку **OK**.



*Рис. 13 – Диалоговое окно Анализ остатков в модуле*

На полученном графике данные с подогнанной прямой имеют вид (рис. 14).



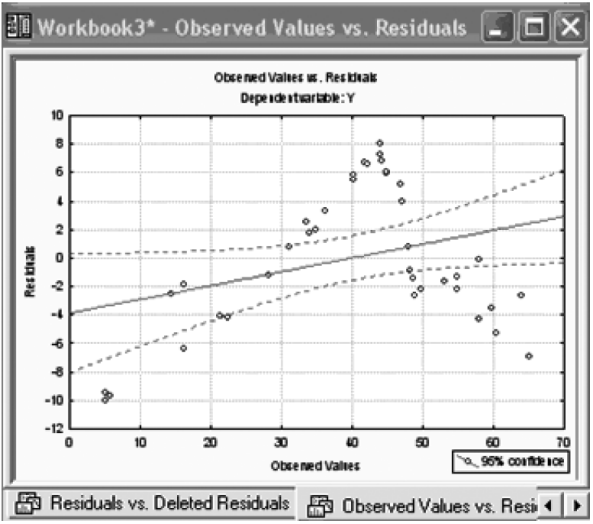
*Рис. 14 – Линейная регрессия для данных X и Y*

**Шаг 6.** Важным элементом анализа является оценка адекватности модели. После того как доказана адекватность модели, полученные результаты можно уверенно использовать для дальнейших действий.

Анализ адекватности основывается на анализе остатков. Остатки представляют собой разности между наблюдаемыми значениями и модельными, то есть значениями, подсчитанными по модели с оцененными параметрами.

В STATISTICA в модуле **Множественная регрессия** имеется специальное диалоговое окно, в котором проводится всесторонний анализ остатков (диалоговое окно **Residual Analysis (анализ остатков)** появится на экране) (рисунок 13).

Нажмите в этом окне кнопку **Obs&residuals**. На экране появится график (рис. 15), который говорит о достаточной адекватности модели.



*Рис. 15 – График наблюдаемых переменных-остатков*