Лабораторная работа № 4

Решение задачи линейного корреляционного и регрессионного анализа

Цель работы – выявить связь между случайными переменными путем оценки коэффициентов корреляции и при установлении этой связи конкретизировать ее, построив регрессионную модель.

1. Теоретический обзор

Коэффициент корреляции ρ_{xy} характеризует тесноту связи между случайными переменными X и Y в генеральной совокупности. Коэффициент корреляции определяется через корреляционный момент (ковариацию) K_{xy} по формуле:

$$\rho_{xy} = \frac{K_{xy}}{\sigma_x \cdot \sigma_y} = \frac{M[(X-m_x)\cdot(Y-m_y)]}{\sigma_x \cdot \sigma_y}. \quad (3.1)$$

Известно, что ρ_{xy} является показателем тесноты связи лишь в случае линейной зависимости между двумя переменными. Для линейно независимых случайных величин $\rho_{xy} \equiv 0$. Но даже и для зависимых СВ ρ_{xy} может быть равен 0. В этом случае СВ X и Y называют *некоррелированными*.

Пусть получена выборка N пар СВ X и Y. Тогда коэффициент корреляции можно оценить по выборочным данным следующим образом:

$$r = \rho_{xy} = \frac{S_{xy}}{S_x S_y}. \quad (3.2)$$

Вспомним "хорошие" (несмещённые, состоятельные и эффективные) оценки:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i; \qquad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i; \quad (3.3)$$

$$S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N-1}\left[\sum_{i=1}^{N} x_i^2 - N\cdot\bar{x}^2\right]; \quad (3.4)$$

$$S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2 = \frac{1}{N-1}\left[\sum_{i=1}^{N} y_i^2 - N\cdot\bar{y}^2\right]; \quad (3.5)$$

$$S_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N-1}\left[\sum_{i=1}^{N} x_i y_i - N\cdot\bar{x}\bar{y}\right]. \quad (3.6)$$

Тогда ***эмпирический коэффициент корреляции*** определяется по формуле:

$$r_{xy} = \frac{\frac{1}{N-1}\left[\sum_{i=1}^{N} x_i y_i - N\cdot\bar{x}\bar{y}\right]}{S_x \cdot S_y} = \frac{\sum_{i=1}^{N} x_i y_i - N\cdot\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{N} x_i^2 - N\bar{x}^2\right)\left(\sum_{i=1}^{N} y_i^2 - N\bar{y}^2\right)}}. \quad (3.7)$$

Как и ρ_{xy} выборочный коэффициент корреляции принимает значения в интервале $[-1;+1]$, причем граничные значения достигаются только при

наличии идеальной линейной связи между наблюдениями. Нелинейная связь и (или) разброс данных, обусловленных неполной коррелированностью СВ или ошибками измерений, приводит к уменьшению абсолютного значения r_{xy}. Эмпирический коэффициент корреляции r_{xy} дает состоятельную, но смещённую оценку. Однако при $N > 50$ величина смещения составляет менее 1%. Для оценки точности выборочного значения r_{xy} удобно использовать некоторую функцию от r_{xy}:

$$W = \frac{1}{2}\ln\left[\frac{1+r_{xy}}{1-r_{xy}}\right]. \qquad (3.8)$$

Распределение случайной величины W можно аппроксимировать нормальным распределением с соответствующим средним и дисперсией:

$$m_W = \frac{1}{2}\ln\left[\frac{1+\rho_{xy}}{1-\rho_{xy}}\right]; \quad \sigma_W^2 = \frac{1}{N-3}. \qquad (3.9)$$

Даже для независимых случайных величин (СВ) эмпирический коэффициент корреляции может быть отличен от "0" вследствие случайного рассеивания результатов измерения. Т.е. из-за выборочной изменчивости необходимо проверять, свидетельствует ли не нулевые значения выборочного коэффициента корреляции о существовании статистически значимой корреляции между исследуемыми СВ X и Y. Сделать это можно, проверив гипотезу H_0: $\rho_{xy} = 0$, причем отклонение гипотезы будет свидетельствовать о принятии альтернативной гипотезы H_1 –*корреляция значимая.*

Из (3.9) следует, что при $\rho_{xy} = 0$ выборочное распределение W будет нормальным со средним $m_w = 0$ и дисперсией $\sigma_W^2 = \frac{1}{N-3}$. Поэтому область принятия гипотезы о нулевой корреляции будет иметь вид:

$$z_{\alpha/2} \le \frac{\sqrt{N-3}}{2}\ln\left[\frac{1+r_{xy}}{1-r_{xy}}\right] \le z_{1-\alpha/2}. \qquad (3.10)$$

Здесь α –уровень значимости, Z –стандартное нормальное распределение $N(0,1)$.

Если корреляционный анализ установит степень взаимосвязи двух и более случайных величин, логичен следующий шаг – построение модели этой связи. Такая модель дала бы возможность предсказать значения одной случайной величины по конкретным значениям другой. А методы решения подобных задач носят название "***регрессионный анализ***".

В линейный регрессионный анализ [3] входит широкий круг задач, связанных с построением (восстановлением) зависимостей между группами числовых переменных $X = (x_1, \ldots x_n)$ и $Y = (y_1, \ldots y_m)$. Предполагается, что X – независимые переменные (факторы, объясняющие переменные) влияют на значения Y – зависимых переменных (откликов, объясняемых переменных). По имеющимся эмпирическим данным (X_i , Y_i), i = 1, ..., n

требуется построить функцию $f(X)$, которая приближенно описывала бы изменение Y при изменении X:

$$Y \approx f(X).$$

Предполагается, что множество допустимых функций, из которого подбирается $f(X)$, является параметрическим:

$$f(X) = f(X, \Theta).$$

Здесь Θ – неизвестный параметр (вообще говоря, многомерный). При построении $f(X)$ будем считать, что

$$Y = f(X, \Theta) + \varepsilon, \qquad (3.11)$$

где первое слагаемое – закономерное изменение Y от X, а второе – ε - случайная составляющая с нулевым средним; $f(X, \Theta)$ является условным математическим ожиданием Y при условии известного X и называется ***регрессией Y по X.***

Пусть X и Y одномерные величины; обозначим их x и y, а функция $f(x, \theta)$ имеет вид $f(x, \theta) = A + bx$, где $\theta = (A\ b)$. Учитывая имеющиеся наблюдения $(x_i, y_i), i = 1, \cdots, n$, полагаем:

$$y_i = A + bx_i + \varepsilon_i, \qquad (3.12)$$

где $\varepsilon_1, \cdots, \varepsilon_n$ – независимые (ненаблюдаемые) одинаково распределенные случайные величины. Можно различными методами подбирать "лучшую" прямую линию. Общепринята такая процедура определения коэффициентов a и b, при которой минимизируется сумма квадратов отклонений наблюденных значений от предсказанных значений. Эта процедура называется ***методом наименьших квадратов*** **(МНК)**.

Построим оценку параметра $\theta = (A, b)$ так, чтобы величины

$$e_i = y_i - f(x_i, \theta) = y_i - A - bx_i,$$

называемые остатками, были как можно меньше, а именно, чтобы сумма их квадратов была минимальной:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i}^{n} (y_i - A - bx_i)^2 = \min \text{ по } (A, b) \qquad (3.13)$$

Чтобы упростить формулы, положим в (3.12) $x_i = x_i - \bar{x} + \bar{x}$, тогда получим:

$$y_i = a + b(x_i - \bar{x}) + \varepsilon_i, \quad i = 1, \cdots, n. \qquad (3.14)$$

Здесь $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $a = A + b\bar{x}$. Минимизируем сумму квадратов отклонений

$$Q = \sum_{i=1}^{N} (y_i - a - b(x_i - \bar{x}))^2, \qquad (3.15)$$

приравняв нулю частные производные по a и b

$$\frac{\partial Q}{\partial a} - \frac{\partial Q}{\partial B} - 0. \qquad (3.16)$$

Полученную систему линейных уравнений решим относительно a и b. Учитывая, что на практике у нас имеется ограниченная выборка из n пар наблюденных значений x и y, решение системы $(\hat{a}, \hat{b})$ легко находится:

$$\hat{a} = \bar{y}, \text{ где } \bar{y} = \sum_{i=1}^{n} y_i/n \tag{3.17}$$

$$\hat{b} = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}. \tag{3.18}$$

Свойства оценок

Нетрудно показать, что если $M[\varepsilon] = 0$, $D[\varepsilon] = \sigma^2$, то

1) $M[\hat{a}] = a$, $M[\hat{b}] = b$, т.е. оценки несмещенные;
2) $D[\hat{a}] = \sigma^2/n$, $D[\hat{b}] = \sigma^2 / \sum_{i=1}^{n}(x_i - \bar{x})^2$
3) $cov(\hat{a}, \hat{b}) = 0$;

Если дополнительно предположить нормальность распределения ε_i, то

4) оценки $\hat{a}$ и $\hat{b}$ нормально распределены и независимы;
5) остаточная сумма квадратов (3.15) независима от $(\hat{a}, \hat{b})$, а величина Q/σ^2 распределена по закону "хи-квадрат" χ^2_{n-2} с $n-2$ степенями свободы.

Оценка для σ^2 и интервальные оценки коэффициентов линейной регрессии

Свойство 5) дает возможность несмещенной оценки неизвестного значения σ^2 величиной

$$S^2 = Q/(n-2). \tag{3.19}$$

Поскольку S^2 независима от $\hat{a}$ и $\hat{b}$, отношения

$$\sqrt{n} \cdot \frac{\hat{a}-a}{S} \quad \text{и} \quad \frac{\hat{b}\ b}{S_b}, \text{ где } S_b = S/\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3.20}$$

имеют распределение Стьюдента с $(n-2)$ степенями свободы. Тогда соответствующие доверительные интервалы (при доверительной вероятности β) будут равны

$$\bar{a} - t_{\frac{\alpha}{2},n-2}\frac{S}{\sqrt{n}} < a < \bar{a} + t_{1-\frac{\alpha}{2},n-2}\frac{S}{\sqrt{n}}$$

$$\bar{b} - t_{\frac{\alpha}{2},n-2}S_b < b < \bar{b} + t_{1-\frac{\alpha}{2},n-2}S_b. \tag{3.21}$$

Здесь $t_{\frac{\alpha}{2},n-2}$ и $t_{1-\frac{\alpha}{2},n-2}$ соответствующие квантили распределения Стьюдента с $n-2$ степенями свободы. Таким образом, найденные интервалы (3.21) с доверительной вероятностью $1-\alpha$ накрывают определяемые параметры (теоретические коэффициенты регрессии).

Проверка гипотез относительно коэффициентов линейной регрессии

На первом этапе регрессионного анализа наиболее важной является задача установления линейной зависимости между переменными y и x. С этой целью сформулируем гипотезы:

H_0: $b = 0$, – линейная зависимость отсутствует, коэффициент угла наклона прямой незначимо отличается от нуля;

H_1: $b \neq 0$, – линейная зависимость значительная и коэффициент угла наклона не равен нулю.

При проверке гипотезы воспользуемся t – статистикой и, если выполняется условие

$$t = \frac{b}{S_b} > t_{1-\frac{\alpha}{2};n-2}, \tag{3.22}$$

то гипотезу H_0 следует отклонить при уровне значимости $\alpha = 1 - \beta$.

Другой способ (в данном случае эквивалентный (3.22)) проверки гипотезы H_0 состоит в вычислении статистики

$$F = \frac{b^2/D[b]}{Q/\sigma^2(n-2)} = \frac{t^2}{S_0^2} \tag{3.23}$$

распределенной, если H_0 верна, по закону $F(1, n-2)$ Фишера с числом степеней свободы 1 и $n-2$. Если

$$F > F(1-\alpha, 1, n-2), \tag{3.24}$$

где $F(1-\alpha, 1, n-2)$ –квантиль уровня $1-\alpha$, то гипотеза H_0 отклоняется с уровнем значимости α.

Аналогичным образом проверяется гипотеза о статистической значимости нулю коэффициента регрессии a (свободный член линейного уравнения равен нулю): $t = \frac{a}{S}$.

Особый интерес представляет выборочное распределение $\hat{y}$ при конкретном значении $x = x_0$. Так как $\hat{y}$ ведет себя как СВ, распределенная по нормальному закону, для нее тоже можно построить доверительный интервал. Соответствующая статистика имеет вид:

$$t_{\hat{y}} = \frac{\hat{y} - y}{S_{y|x}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right)^{\frac{1}{2}}}. \tag{3.25}$$

В выражении (3.25) величина $S_{y|x}$ – это выборочное стандартное отклонение наблюденного значения y_i от предсказанного $\hat{y} = a + b(x_i - \bar{x})$, равное

$$S_{y|x} = \left[\frac{\sum(y_i - \hat{y}_i)^2}{n-2}\right]^{\frac{1}{2}} = \left[\frac{n-1}{n-2} \cdot S_y^2(1 - r_{xy}^2)\right]^{\frac{1}{2}}. \tag{3.26}$$

Проверка качества уравнения регрессии

Оценим, насколько хорошо модель линейной регрессии описывает данную систему наблюдений. В качестве этой оценки воспользуемся коэффициентом детерминации.

Рассмотрим следующие вариации (суммы квадратов отклонений):

$T_{ss} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ –(*total sum of square*) разброс фактических значений от их среднего арифметического;

$R_{ss} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ – (*regression sum of square*) разброс обусловленный регрессией от их среднего арифметического;

$E_{ss} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ – (*error sum of square*) разброс за счет случайных отклонений от функции регрессии.

Оказывается,

$$T_{SS} = R_{SS} + E_{SS}, \quad (3.27)$$

т.е. полный разброс равен сумме разбросов за счет регрессии и за счет случайных отклонений. Величина R_{SS}/T_{SS} – это доля вариации значений y_i, обусловленной регрессией (т.е. доля закономерной изменчивости в общей изменчивости). Статистика

$$R^2 = R_{SS}/T_{SS} = 1 - E_{SS}/T_{SS} \quad (3.28)$$

называется ***коэффициентом детерминации***.

При $R^2 = 0$ регрессия ничего не дает, т.е. знание x не улучшает предсказания для y по сравнению с тривиальным $\hat{y}_i = \bar{y}$. Другой крайний случай $R^2 = 1$ означает точную подгонку: все точки наблюдений лежат на регрессионной прямой. Чем ближе к 1 значение R^2, тем лучше качество подгонки (регрессионной модели).

2. Корреляционный и регрессионный анализ в пакете Statistica 6.0

Анализ проведем с данными, представленными в файле **Product. sta.** Приведенные наблюдения по 45 предприятиям легкой промышленности, отражают статистические связи между стоимостью основных фондов (*Fonds,* млн руб.) и средней выработкой на 1 работника (*Product,* тыс. руб.). Также представлен вспомогательный признак – z: z = 1 – предприятие федерального подчинения, z = 2 – муниципальное.

Для построения корреляционной матрицы воспользуемся модулем: *Statistics* → *Basic Statistic and Tables* → *Correlation matrices*. Выберем все переменные и нажмем кнопку Summary.

Матрица коэффициентов корреляции – симметрична относительно главной диагонали. Значения на главной диагонали равны 1 и не указывают на строгую линейную зависимость, т.к. они получены при делении i – й дисперсии на саму себя. Остальные значения могут быть либо близки к нулю (менее 0.1), либо отличаются от 0. Возникает вопрос, эти флуктуации обусловлены статистикой выборки, либо переменные действительно коррелированны? Эту задачу можно корректно решить в рамках проверки гипотезы о равенстве нулю эмпирического коэффициента корреляции (гипотеза H_0), если коэффициент корреляции значимо отличается от нуля, то принимается альтернативная гипотеза – переменные коррелированны.

В приведенной таблице коэффициенты, значимо отличные от ноля выделены **красным**. Все оценки проведены для уровня значимости $\alpha = 0.05$.

Для получения более подробной информации в закладке *Options* отметим пункт Display r, p-levels.

Теперь матрица коэффициентов примет вид:

Variable	Correlations (Product.sta) Marked correlations are s N=45 (Casewise deletion		
	Fonds	Product	Z
Fonds	1,0000	,7723	-,1371
	p= ---	p=,000	p=,369
Product	,7723	1,0000	-,2084
	p=,000	p= ---	p=,170
Z	-,1371	-,2084	1,0000
	p=,369	p=,170	p= ---

Рис. 3.1. Результаты линейного корреляционного анализа

Анализ результатов свидетельствует, что для переменных Fonds и Product коэффициент корреляции незначимо отличается от 0 с вероятностью ~ 0.0001, т.е. принимаем альтернативную гипотезу – переменные коррелируют.

Убедимся, что предположение о линейной зависимости переменных не лишено смысла, для чего предварительно построим диаграмму рассеяния, выполнив последовательно действия *Graphs - 2D Graphs - Scatter plots - Variables - X: Fonds, Y : Product, Advanced, Graphs Type: Regular, Fit* (подбор): *Linear , Elipse, Normal coefficient 0.95- OK - OK* (см. рис. 3.2).

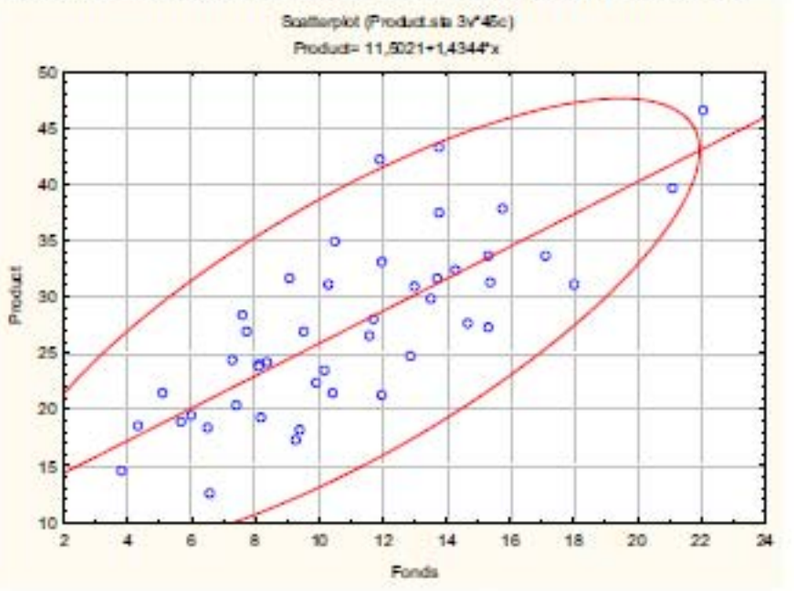

Рис. 3.2. Диаграмма рассеяния с подобранной прямой регрессии

Диаграмма рассеяния с наблюдениями, вытянутыми вдоль линии регрессии подтверждает наши предположения. На следующем этапе приступим к количественному анализу, используя при этом модуль ***Multiple Regression*** (множественная регрессия).

В стартовом диалоговом окне этого модуля при помощи кнопки *Variables* указываем зависимую переменную *Dependent var:* *Product* и независимую *Independent var:* *Fonds* - *OK*. В поле *Input file* указывается тип файла с данными *Raw Data* – данные в виде строчной таблицы. В поле *MD deletion* указываем способ исключения из обработки недостающих данных –

Casewise (игнорируется вся строка, в которой есть хотя бы одно пропущенное значение).

После выбора всех опций стартового диалогового окна регрессионного анализа и нажатия кнопки *OK* появляется окно результатов регрессионного анализа *Multiple Regressions Results.* Прежде, чем анализировать полученные результаты, опишем наиболее важные параметры полученной регрессионной модели:

- *Multiple R* – коэффициент множественной корреляции, характеризующий тесноту линейной связи между зависимой и всеми независимыми переменными;
- R^2 – коэффициент детерминации, выражающий долю вариации зависимой переменной, объясненную с помощью регрессионного уравнения;
- *adjusted R* – скорректированный коэффициент множественной корреляции. Включение новой переменной в регрессионное уравнение увеличивает R^2 не всегда, а только в том случае, когда частный *F*–критерий при проверке гипотезы о значимости включаемой переменной больше или равен 1. В противном случае включение новой переменной уменьшает значение R^2 и *adjusted R*;
- *F* – критерий используется для проверки значимости регрессии (в качестве нулевой гипотезы проверяется гипотеза – между зависимой и независимыми переменными нет линейной зависимости);
- *df* – числа степеней свободы для *F*– критерия;
- *p* – вероятность нулевой гипотезы для *F*– критерия;
- *Standard error of estimate* – стандартная ошибка оценки (уравнения); Эта оценка является мерой рассеяния наблюденных значений относительно регрессионной прямой;
- *Intercept* – оценка свободного члена уравнения;

- *Std.Error* – стандартная ошибка оценки свободного члена уравнения;
- *t* – критерий для оценки свободного члена уравнения;
- *p* – вероятность нулевой гипотезы для свободного члена уравнения.
- *Beta* – β – коэффициенты уравнения. Это стандартизированные регрессионные коэффициенты, рассчитанные по стандартизированным значениям переменных. По их величине можно оценить значимость зависимых переменных. Коэффициент показывает, на сколько единиц стандартного отклонения изменится зависимая переменная при изменении на одно стандартное отклонение независимой переменной, при условии постоянства остальных независимых переменных. Свободный член в таком уравнении равен 0.

В окне *Multiple Regression Results* получили такие результаты: коэффициент детерминации $R^2 = 0.597$; гипотеза о нулевом значении наклона отклоняется с высоким уровнем значимости $p = 0.000000$ (т.е. $p < 10^{-6}$). Нажмем кнопку *Regression summary* – получим таблицу результатов (рис. 3.3).

	Regression Summary for Dependent Variable: Product (Product.sta) R= ,77227708 R?= ,59641189 Adjusted R?= ,58702612 F(1,43)=63,544 p<,00000 Std.Error of estimate: 5,0082					
N=45	**Beta**	**Std. Err. of Beta**	B	**Std. Err. of B**	**t(43)**	**p-Level**
Intercept			11,50212	**2,128204**	**5,404612**	**0,000003**
Fonds	0,772277	0,096880	1,43440	0,179942	7,971466	0,000000

Рис. 3.3. Результаты линейного регрессионного анализа

В ее заголовке повторены результаты предыдущего окна; в столбцах приведены: *B* – значения оценок неизвестных коэффициентов регрессии; *St. Err. of B* – стандартные ошибки оценки коэффициентов, *t* – значение статистики Стьюдента для проверки гипотезы о нулевом значении коэффициента; *p – level* – уровень значимости принятия этой гипотезы. В данном случае, поскольку значения *p-level* очень малы (меньше 10^{-5}), гипотезы о нулевых значениях коэффициентов отклоняются с высоким уровнем значимости. Итак, линейная модель имеет вид:

$$\textbf{\textit{Product}} = 11.5 + 1.43\ \textbf{\textit{Fonds}}.$$

Соответствующие стандартные ошибки коэффициентов равны: 2.1 и 0.18. Значение коэффициента детерминации $R^2 = 0.597$ достаточно велико ($R = 0.77$, т.е. 77 % всей изменчивости объясняется вариацией фондов).

Было бы логично предположить, что при более однородной совокупности предприятий – для предприятий федерального подчинения (z=1) регрессионная модель окажется более качественной. Предварительно визуально оценим данные процедурой *Scatterplot* (при отборе наблюдений используем кнопку *Select cases→Use selection conditions for this Analysis/Graph only→Include cases→Specific, selected by:→By Expression:* z=1. Сравнивая диаграммы рассеяния рис.3.2 и рис. 3.4 видим, что эллипс рассеяния более вытянут вдоль регрессионной прямой, причем все наблюдения находятся внутри эллипса.

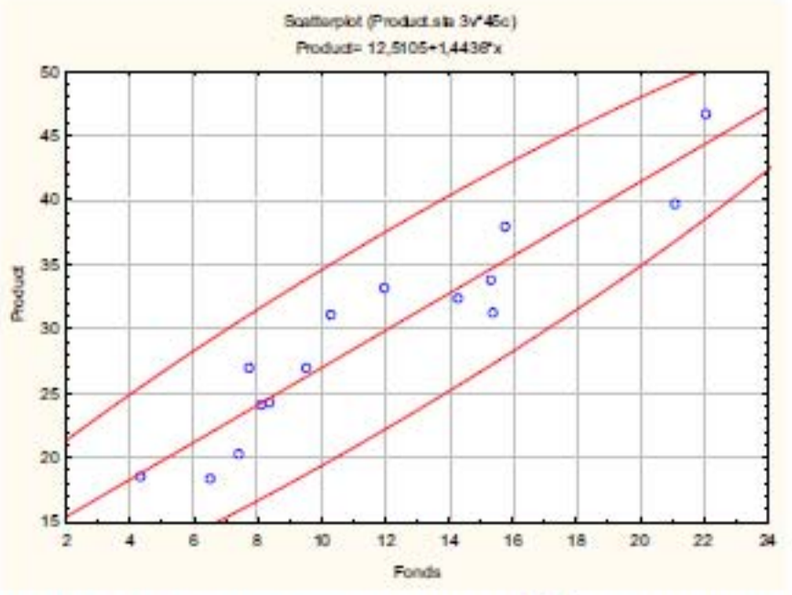

Рис. 3.4. Диаграмма рассеяния для предприятий федерального подчинения

Возвращаемся в окно *Multiple Regression* - *Select cases* - в окне *Case Selection Conditions* (условия выбора наблюдений $z = 1$) - *OK* - *OK* - в окнах *M.R.Results* и *Regression summary* получаем результаты:

N=15	Regression Summary for Dependent Variable: Product (Product.sta) R= ,94717253 R?= ,89713581 Adjusted R?= ,88922318 F(1,13)=113,38 p<,00000 Std.Error of estimate: 2,6886					
	Beta	**Std. Err. of Beta**	B	**Std. Err. of B**	**t(43)**	**p-Level**
Intercept			12,51054	**1,753810**	**7,13335**	**0.000008**
Fonds	0,947173	0,088953	1,44356	0,135571	10,64802	0.000000

Рис. 3.5. Регрессионный анализ предприятий федерального подчинения

Теперь линейная модель примет вид:

$$\textbf{\textit{Product}} = 12.51 + 1.44\ \textbf{\textit{Fonds}}.$$

Коэффициент детерминации увеличился с 0.597 до 0.897, значения остальных параметров тоже улучшились (ошибки уменьшились).

Для расчета по полученному регрессионному уравнению значений зависимой переменной (***Product***) по значениям независимой переменной (***Fonds***) воспользуемся кнопкой 

(раздел Residuals/assumptions/prediction). Зададим значение ***Fonds*** = 18, и учтем, что в пакете Statistica приводится как точечная, так и интервальная оценка (см. рис. 3.6).

Variable	Predicting Values for (Product.sta) variable: Product Include condition: z=1		
	B-Weight	Value	B-Weight * Value
Fonds	1,443557	18,00000	25,98403
Intersept			12,51054
Predicted			38,49457
-95,0%CL			36,15750
+95,0%CL			40,83164

Рис. 3.6. Предсказанные точечные и интервальные оценки зависимой переменной

Анализ остатков

Остатки – это разности между опытными и предсказанными значениями зависимой переменной в построенной регрессионной модели.

Кнопка *Perform residual analysis* в модуле *Residuals/assumptions/prediction* запускает процедуру всестороннего анализа остатков регрессионного уравнения (см. рис. 3.7). При анализе остатков следует учитывать ряд существенных факторов:

- Если модель подобрана правильно, то остатки (столбец Residuals в *Predicted & Residuals Values*) будут вести себя достаточно хаотично, в известном смысле они будут напоминать белый шум.
- В остатках не будет систематической составляющей, резких выбросов, в чередовании их знаков не будет никаких закономерностей, остатки будут независимы друг от друга.

При анализе остатков весьма полезной характеристикой является расстояние Махаланобиса (Mahalanobis Distance). Независимые переменные в уравнении регрессии можно представлять точками в многомерном пространстве (каждое наблюдение изображается точкой). В этом пространстве можно построить точку центра (среднюю точку). Эта "средняя точка" в многомерном пространстве называется центроидом, т.е. центром тяжести. Расстояние Махаланобиса определяется как расстояние от наблюдаемой точки до центра тяжести в многомерном пространстве. Соответственно, значения расстояния Махаланобиса, которые достаточно отличаются от остальных, указывают на выбросы.

	Predicted & Residual Values (Product.sta) Dependent variable: Product Include condition: z=1								
Case No.	Observed Value	Predicted Value	Residual	Standard Pred. v.	Standard Residual	Std.Err. Pred.Val	Mahalanobis Distance	Deleted Residual	Cook's Distance
1	18,30000	21,89366	-3,59366	-1,01506	-1,33665	1,006910	1,030354	-4,17996	0,169520
2	31,10000	27,37918	3,72082	-0,29810	1,38395	0,726478	0,088866	4,01389	0,081371
3	27,00000	23,62593	3,37407	-0,78866	1,25498	0,896114	0,621977	3,79575	0,110718
4	37,90000	35,31874	2,58126	0,73960	0,96009	0,874250	0,547008	2,88647	0,060940
5	20,30000	23,19286	-2,89286	-0,84526	-1,07599	0,922371	0,714460	-3,27877	0,087525
6	32,40000	33,15341	-0,75341	0,45659	-0,28023	0,767805	0,208474	-0,82031	0,003796
7	31,20000	34,74132	-3,54132	0,66413	-1,31719	0,842387	0,441070	-3,92682	0,104713
8	39,70000	42,96960	-3,26960	1,73957	-1,21612	1,429786	3,026102	-4,55894	0,406599
9	46,60000	44,41315	2,18684	1,92824	0,81339	1,549704	3,718121	3,27493	0,246489
10	33,10000	29,83323	3,26677	0,02264	1,21507	0,694372	0,000513	3,50025	0,056530
11	26,90000	26,22433	0,67567	-0,44904	0,25131	0,765504	0,201640	0,73528	0,003032
12	24,00000	24,20335	-0,20335	-0,71319	-0,07564	0,862844	0,508634	-0,22670	0,000366
13	24,20000	24,63642	-0,43642	-0,65658	-0,16233	0,839327	0,431102	-0,48355	0,001576
14	33,70000	34,59697	-0,89697	0,64526	-0,33362	0,834781	0,416365	-0,99266	0,006571
15	18,50000	18,71784	-0,21784	-1,43015	-0,08102	1,240121	2,045315	-0,27671	0,001127
Minimum	18,30000	18,71784	-3,59366	-1,43015	-1,33665	0,694372	0,000513	-4,55894	0,000366
Maximum	46,60000	44,41315	3,72082	1,92824	1,38395	1,549704	3,718121	4,01389	0,406599
Mean	29,66000	29,66000	0,00000	0,00000	0,00000	0,950184	0,933333	-0,03586	0,089392

Рис. 3.7. Анализ остатков регрессионной модели

Для наглядного анализа поведения остатков построим их на нормальной вероятностной бумаге (*Normal plot of residuals*) рис. 3.8.

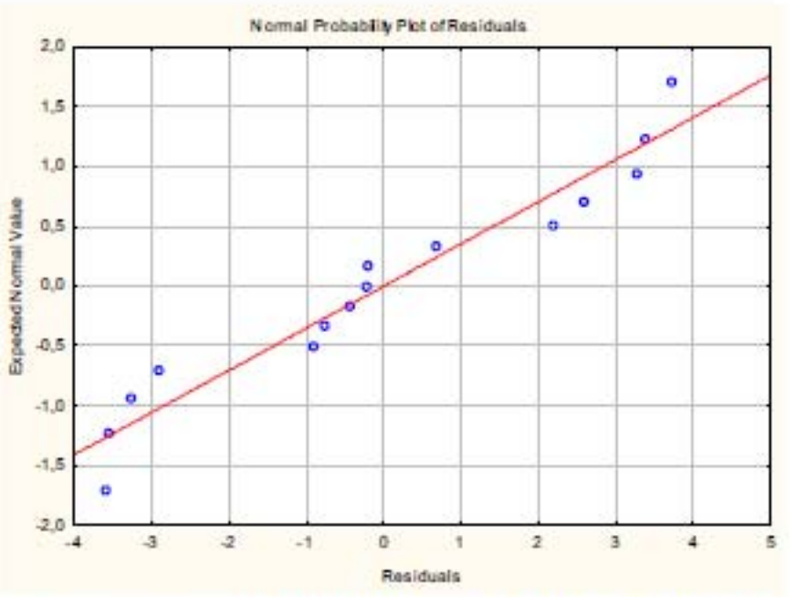

Рис. 3.8. Анализ остатков на нормальной вероятностной бумаге

Отсутствие больших отклонений и группирование остатков вдоль прямой свидетельствует о высоком качестве модели.

3. Задание

Необходимо исследовать зависимость урожайности y зерновых культур (ц/га) от ряда факторов (переменных) сельскохозяйственного производства [4], а именно:

x_1 - число тракторов на 100 га;

x_2 - число зерноуборочных комбайнов на 100 га;

x_3 - число орудий поверхностной обработки почвы на 100 га;

x_4 - количество удобрений, расходуемых на гектар (т/га);

x_5 - количество химических средств защиты растений, расходуемых на гектар (ц/га).

Исходные данные для 20 районов области приведены в табл. 3.1.

Таблица 3.1

	y	x_1	x_2	x_3	x_4	x_5
1	9.7	1.59	.26	2.05	.32	.14
2	8.4	.34	.28	.46	.59	.66
3	9.0	2.53	.31	2.46	.30	.31
4	9.9	4.63	.40	6.44	.43	.59
5	9.6	2.16	.26	2.16	.39	.16
6	8.6	2.16	.30	2.69	.32	.17
7	12.5	.68	.29	.73	.42	.23
8	7.6	.35	.26	.42	.21	.08
9	6.9	.52	.24	.49	.20	.08
10	13.5	3.42	.31	3.02	1.37	.73
11	9.7	1.78	.30	3.19	.73	.17
12	10.7	2.40	.32	3.30	.25	.14
13	12.1	9.36	.40	11.51	.39	.38
14	9.7	1.72	.28	2.26	.82	.17
15	7.0	.59	.29	.60	.13	.35
16	7.2	.28	.26	.30	.09	.15
17	8.2	1.64	.29	1.44	.20	.08
18	8.4	.09	.22	.05	.43	.20

19	13.1	.08	.25	.03	.73	.20
20	8.7	1.36	.26	.17	.99	.42

1. Создайте таблицу $5v \times 20c$. В первый столбец занесите значения переменной y, в остальные столбцы занесите данные, соответствующие вашему варианту (см. табл. 3.2).
2. Постройте матрицу коэффициентов парных корреляции и проанализируйте полученные результаты.
3. Создайте одномерную модель линейной регрессии, выбирая в качестве аргумента (независимой переменной) переменные, находящиеся в столбцах 2 – 5. Оцените качество полученных моделей, обоснуйте выбор лучшей.

Таблица 3.2

N	*II*	*III*	*IV*	*V*
1	x_1	x_2	$(x_1)^2$	$(x_2)^2$
2	x_1	x_3	x_1x_3	$(x_3)^2$
3	x_1	x_4	$x_1 \times x_4$	$(x_4)^2$
4	x_1	x_5	x_5x_1	$\ln(x_5)$
5	x_2	x_3	$\exp(x_2)$	$\exp(x_3)$
6	x_2	x_4	$\ln(x_4)$	$x_2 \times x_4$
7	x_2	x_5	x_2+x_5	$y \times x_5$
8	x_3	x_4	x_3x_4	$\exp(x_4)$
9	x_3	x_5	x_5-x_3	$\ln(x_3)$
10	x_4	x_5	$y \times x_4$	$(x_5)^2$
11	x_1	x_2	$\ln(x_1)$	$\ln(x_2)$
12	x_1	x_3	$(x_1)^2$	x_3x_1
13	x_1	x_4	x_1/x_4	x_4/ x_1
14	x_1	x_5	$x_5 \times x_1$	$\exp(x_5)$
15	x_2	x_3	$\exp(x_2)$	$x_2 \times x_3$
16	x_2	x_4	$\ln(x_2)$	x_2/x_4
17	x_2	x_5	x_2-x_5	$y \times x_5$
18	x_3	x_4	x_4x_3	$\exp(y \times x_4)$
19	x_3	x_5	$\ln(x_5-x_3)$	$y \times x_3$
20	x_4	x_5	$y \times x_5$	$\ln(x_5)^2$
21	x_1	x_2	$\exp(x_1)$	$\exp(x_2)$
22	x_1	x_3	x_3x_1	$\ln(x_3)$
23	x_1	x_4	$\ln(x_1 \times x_4)$	$\exp(x_4)$
24	x_1	x_5	$(x_5)^2$	$\exp(x_1)$
25	x_2	x_3	$y \times x_2$	$\ln(x_2 \times x_3)$
26	x_2	x_4	$\ln(x_2 \times x_4)$	x_2+x_4

4. Контрольные вопросы

1. С какой целью в линейном регрессионном анализе применяется метод наименьших квадратов?
2. Как изменится доверительный интервал для выборочного значения $\hat{y}$ при различных значениях аргумента x?
3. Что такое коэффициент детерминации?
4. Опишите процедуру проверки гипотез относительно коэффициентов линейной регрессии. Кстати, какое распределение имеет используемая при этом статистика?

5. Как проверить значимость (качество) уравнения регрессии? Какая статистика используется при этом?

6. Проанализируйте остатки в регрессионной модели. Каким требования (остаткам) отвечает качественная регрессионная модель?

7. Какой вид графика остатков на нормальной вероятностной бумаге подтвердит качество регрессионной модели?

Список литературы

1 Руководство пользователя пакета Statistica 5.1: http://exponenta.ru/soft/Statist/stat5_1/1/1.asp

2 Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников.- М.: ФИЗМАТЛИТ, 2006,-816 с.

3 Горицкий Ю.А. Практикум по статистике с пакетом Statistica (часть 2): http://exponenta.ru/educat/systemat/goritskii/part2/lr.asp

4 Магнус Я.Р., Катышев П.К., Пересецкий А.А. Эконометрика. Начальный курс. М.: Дело, 1998. 248с.