**Министерство науки и высшего образования Российской Федерации**

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ "МЭИ"**

Институт информационных и вычислительных технологий

Кафедра математического и компьютерного моделирования

# Отчёт по лабораторной работе №8
# "Регрессионный анализ"

Студент: Симаков А.М.
Преподаватель: Шевченко О.В.

Москва 2023

# 1 Введение

В линейный регрессионный анализ входит широкий круг задач, связанных с построением (восстановлением) зависимостей между группами числовых переменных

$$X \equiv (x_1, ..., x_p), \quad Y = (y_1, ..., y_m)$$

Предполагается, что $X$ - независимые переменные (факторы, объясняющие переменные) влияют на значения $Y$ - зависимых переменных (откликов, объясняемых переменных). По имеющимся эмпирическим данным $(X_i, Y_i)$, $i \in [1, n] \cap \mathbb{N}$ 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:

$$f = f(X, \theta),$$

где $\theta$ - неизвестный параметр (вообще говоря, многомерный). При построении $f(X)$ будем считать, что

$$Y = f(X, \theta) + \varepsilon,$$

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

# 2 Простая линейная регрессия

Пусть $X$ и $Y$ одномерные величины; обозначим их $x$ и $y$, а функция $f(x, \theta)$ имеет вид $f(x, \theta) = A + bx$, где $\theta = (A, b)$. Относительно имеющихся наблюдений $(x_i, y_i)$, $i \in [1, n] \cap \mathbb{N}$, полагаем, что

$$y_i = A + bx_i + \varepsilon_i,$$

где $\varepsilon_1, ..., \varepsilon_n$ - независимые (ненаблюдаемые) одинаково распределенные случайные величины. Можно различными методами подбирать "лучшую" прямую линию. Широко используется **метод наименьших квадратов**.

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:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - A - bx_i)^2 \to \min_{(A,b)}$$

Чтобы упростить формулы, положим $x_i = x_i - \overline{x} + \overline{x}$. Тогда

$$y_i = a + b(x_i - \overline{x}) + \varepsilon_i, \quad i \in [1, n] \cap \mathbb{N},$$

где $\overline{x} = \sum_{i=1}^{n} x_i/n, \quad a = A + b\overline{x}$
Сумму

$$\sum_{i=1}^{n} (y_i - a - b(x_i - \overline{x}))^2$$

минимизируем по $(a, b)$, приравнивая нулю производные по $a$ и $b$. Получим систему линейных уравнений относительно $a$ и $b$. Ее решение $(\widehat{a}, \widehat{b})$ легко находится

$$\widehat{a} = \overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$
$$\widehat{b} = \frac{\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

**Свойства оценок**

Нетрудно показать, что если $M\varepsilon_i = 0, \quad D\varepsilon_i = \sigma^2$, то
1. $M\widehat{a} = a, \quad M\widehat{b} = b \implies$ обе оценки несмещённые
2. $D\widehat{a} = \sigma^2/n, \quad D\widehat{b} = \sigma^2 / \sum_{i=1}^{n} (x_i - \overline{x})^2$
3. $cov(\widehat{a}, \widehat{b}) = 0$
Если дополнительно предположить нормальность распределения $\varepsilon_i$ , то
4. $\widehat{a}, \widehat{b}$ имеют нормальное распределение и независимы
5. Остаточная сумма квадратов

$$Q^2 = \sum_{i=1}^{n} \left( y_i - \widehat{a} - \widehat{b}(x_i - \overline{x}) \right)^2$$

независима от $(\widehat{a}, \widehat{b})$, а $Q^2/\sigma^2 \sim \chi^2_{n-2}$

### Оценка для $\sigma^2$ и доверительные интервалы

Свойство 5 даёт возможность несмещенно оценивать неизвестный параметр $\sigma^2$ величиной

$$s^2 = \frac{Q^2}{n-2}$$

Поскольку $s^2$ независима от $\widehat{a}$ и $\widehat{b}$, отношения

$$\sqrt{n}\frac{\widehat{a}-a}{s}, \quad \frac{\widehat{b}-b}{s_b}: \quad s_b = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}}$$

имеют распределение Стьюдента с $(n-2)$ степенями свободы, и потому доверительные интервалы для $a$ и $b$ таковы:

$$|\widehat{a}-a| \le t_p\frac{s}{\sqrt{n}}, \quad |\widehat{b}-b| \le t_p s_b,$$

где $tp$ - квантиль уровня $(1+P_d)/2$ распределения Стьюдента с $n-2$ степенями свободы, $P_d$ - коэффициент доверия.

### Проверка гипотезы о коэффициенте наклона

Обычно возникает вопрос: может быть, $y$ не зависит от $x$, т.е. $b=0$, и изменчивость $y$ обусловлена только случайными составляющими $\varepsilon_i$? Проверим гипотезу $H: b=0$. Если 0 не входит в доверительный интервал для $b$, т.е.

$$\frac{|b|}{s_b} > t_p,$$

то гипотезу $H$ следует отклонить. Уровень значимости при этом $\alpha = 1-P_d$.

Другой способ (в данном случае эквивалентный) проверки гипотезы $H$ состоит в вычислении статистики

$$F = \frac{\widehat{b}^2/D\widehat{b}}{Q^2/(\sigma^2(n-2))} = \frac{\widehat{b}^2}{s_b^2},$$

распределенной, если $H$ верна, по закону $\mathcal{F}(1,n-2)$ Фишера с числом степеней свободы 1 и $n-2$. Если

$$F > F_{1-\alpha},$$

где $F_{1-\alpha}$ - квантиль уровня $1-\alpha$ распределения $\mathcal{F}(1,n-2)$, то гипотеза $H$ отклоняется с уровнем значимости $\alpha$.

### Вариация зависимой переменной и коэффициент детерминации

Рассмотрим вариацию (разброс) $T_{ss}$ (*total sum of square*) значений $y_i$ относительно среднего значения $\overline{y}$

$$T_{ss} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

Обозначим $\widehat{y_i}$ предсказанные с помощью функции регрессии значения $y_i$ : $\widehat{y} = \widehat{a} + \widehat{b}x_i$. Сумма $R_{ss}$ (*regression sum of square*)

$$R_{ss} = \sum_{i=1}^{n} (\widehat{y_i} - \overline{y})^2$$

означает величину разброса, которая обусловлена регрессией (ненулевым значением наклона $\widehat{b}$). Сумма $E_{ss}$ (*error sum of squares*)

$$E_{ss} = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2$$

означает разброс за счет случайных отклонений от функции регрессии.
Оказывается,

$$T_{ss} = R_{ss} + E_{ss},$$

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

$$R^2 = \frac{R_{ss}}{T_{ss}} = 1 - \frac{E_{ss}}{T_{ss}}$$

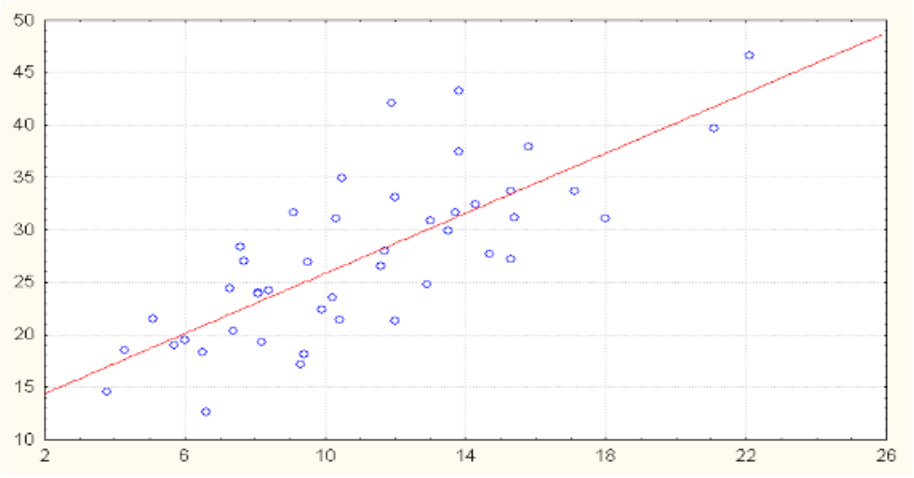
называется **коэффициентом детерминации**. Если $R^2 = 0$, это означает, что регрессия ничего не дает, т.е. знание $x$ не улучшает предсказания для $y$ по сравнению с тривиальным $\widehat{y_i} = \overline{y}$. Другой крайний случай $R^2 = 1$ означает точную подгонку: все точки наблюдений лежат на регрессионной прямой. Чем ближе к 1 значение $R^2$ , тем лучше качество подгонки.

# 3 Пример

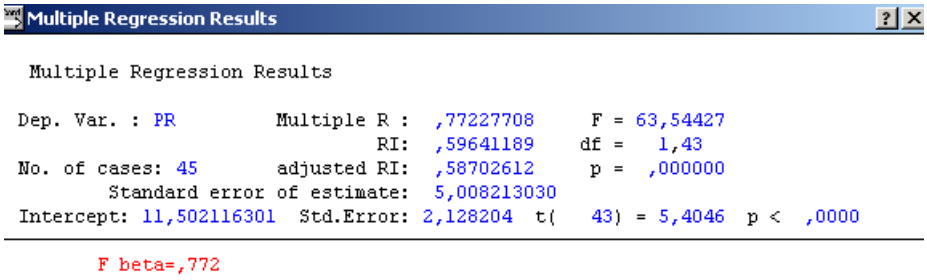
В таблице приведены данные по 45 предприятиям легкой промышленности по статистической связи между стоимостью основных фондов (*fonds*, млн руб.) и средней выработкой на 1 работника (*product*, тыс. руб.); $z$ - вспомогательный признак: $z = 1$ - федеральное подчинение, $z = 2$ – муниципальное.
Таблица

| *fonds* | *product* | *z* | *fonds* | *product* | *z* | *fonds* | *product* | *z* |
|---|---|---|---|---|---|---|---|---|
| 6,5 | 18,3 | 1 | 9,3 | 17,2 | 2 | 10,4 | 21,4 | 2 |
| 10,3 | 31,1 | 1 | 5,7 | 19,0 | 2 | 10,2 | 23,5 | 2 |
| 7,7 | 27,0 | 1 | 12,9 | 24,8 | 2 | 18,0 | 31,1 | 2 |
| 15,8 | 37,9 | 1 | 5,1 | 21,5 | 2 | 13,8 | 43,2 | 2 |
| 7,4 | 20,3 | 1 | 3,8 | 14,5 | 2 | 6,0 | 19,5 | 2 |
| 14,3 | 32,4 | 1 | 17,1 | 33,7 | 2 | 11,9 | 42,1 | 2 |
| 15,4 | 31,2 | 1 | 8,2 | 19,3 | 2 | 9,4 | 18,1 | 2 |
| 21,1 | 39,7 | 1 | 8,1 | 23,9 | 2 | 13,7 | 31,6 | 2 |
| 22,1 | 46,6 | 1 | 11,7 | 28,0 | 2 | 12,0 | 21,3 | 2 |
| 12,0 | 33,1 | 1 | 13,0 | 30,9 | 2 | 11,6 | 26,5 | 2 |
| 9,5 | 26,9 | 1 | 15,3 | 27,2 | 2 | 9,1 | 31,6 | 2 |
| 8,1 | 24,0 | 1 | 13,5 | 29,9 | 2 | 6,6 | 12,6 | 2 |
| 8,4 | 24,2 | 1 | 10,5 | 34,9 | 2 | 7,6 | 28,4 | 2 |
| 15,3 | 33,7 | 1 | 7,3 | 24,4 | 2 | 9,9 | 22,4 | 2 |
| 4,3 | 18,5 | 1 | 13,8 | 37,4 | 2 | 14,7 | 27,7 | 2 |

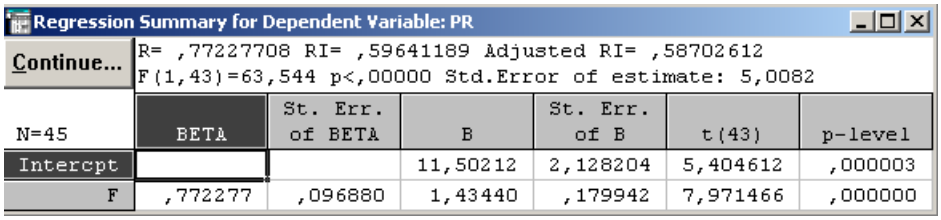
Построим диаграмму рассеяния наблюдений, чтобы убедиться, что предположение линейности регрессионной зависимости не лишено смысла.



Теперь выполним регрессионный анализ.




В окне *Multiple Regression Results* имеем основные результаты: коэффициент детерминации $R^2 = RI$ : 0.596. Гипотеза о нулевом значении наклона отклоняется с высоким уровнем значимости $p = 0.000000$ (т.е. p $< 10^{-6}$). Кнопка *Regression summary* – на экране таблица результатов:




| N=45 | BETA | St. Err. of BETA | B | St. Err. of B | t(43) | p-level |
|---|---|---|---|---|---|---|
| Intercpt | | | 11,50212 | 2,128204 | 5,404612 | ,000003 |
| F | ,772277 | ,096880 | 1,43440 | ,179942 | 7,971466 | ,000000 |

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 отклонения этой гипотезы. В данном случае, поскольку значения $p$-*level* очень малы (меньше $10^{-4}$), гипотезы о нулевых значениях коэффициентов отклоняются с высокой значимостью. Итак, имеем регрессию:

$$product = 11.5 + 1.43 fonds,$$

соответствующие стандартные ошибки коэффициентов: 2.1 и 0.18; значение $s$ : $s = 5.01$ (*Std Error of estimate* - ошибка прогноза выработки по фондам с помощью этой функции). Значение коэффициента детерминации $R^2 = RI = 0.596$ достаточно велико (доля $R = 0.77$ всей изменчивости объясняется вариацией фондов).

Уравнение регрессии показывает, что увеличение основных фондов на 1 млн руб. приводит к увеличению выработки 1 работника в среднем на $\beta_1 = 1.43$ тыс. руб.

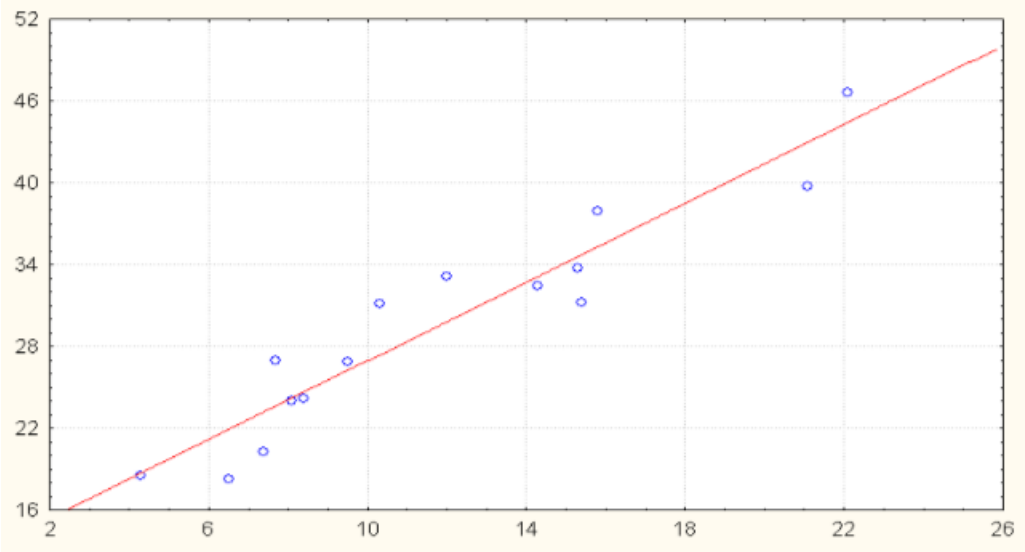
Для удобства интерпретации параметра $\beta_1 = \Delta y/\Delta x$ пользуются коэффициентом эластичности

$$Э = \beta_1 \frac{\overline{x}}{\overline{y}} = \frac{\Delta y}{\Delta x} \cdot \frac{\overline{x}}{\overline{y}} = \frac{\Delta y}{\overline{y}} \Big/ \frac{\Delta x}{\overline{x}},$$

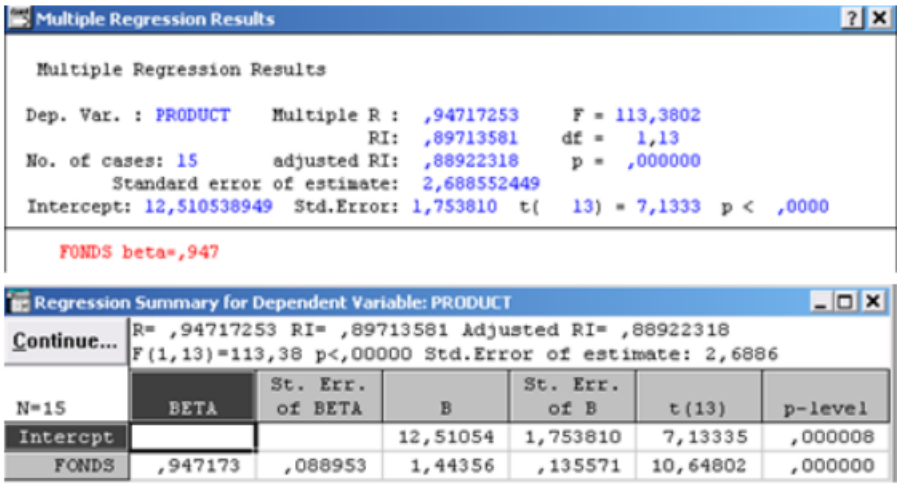
который показывает среднее изменение (в долях или %) зависимой переменной $y$ при изменении фактора $x$:

$$\frac{\Delta y}{\overline{y}} = Э \, \frac{\Delta x}{\overline{x}}$$

Теперь построим регрессию выработки по фондам для более однородной совокупности - для предприятий федерального подчинения ($z = 1$). Можно ожидать, что качество подгонки улучшится. Предварительно визуально оценим данные процедурой *Scatterplot*.



Возвращаемся в окно *Multiple Regression*. В окнах *M.R.Results* и *Regression summary* получаем результаты:



Multiple Regression Results

Multiple Regression Results

Dep. Var. : PRODUCT  Multiple R : ,94717253  F = 113,3802
RI: ,89713581  df = 1,13
No. of cases: 15  adjusted RI: ,88922318  p = ,000000
Standard error of estimate: 2,688552449
Intercept: 12,510538949  Std.Error: 1,753810  t( 13) = 7,1333  p < ,0000

FONDS beta=,947

Regression Summary for Dependent Variable: PRODUCT

Continue...

R= ,94717253 RI= ,89713581 Adjusted RI= ,88922318
F(1,13)=113,38 p<,00000 Std.Error of estimate: 2,6886

| N=15 | BETA | St. Err. of BETA | B | St. Err. of B | t(13) | p-level |
|---|---|---|---|---|---|---|
| Intercpt | | | 12,51054 | 1,753810 | 7,13335 | ,000008 |
| FONDS | ,947173 | ,088953 | 1,44356 | ,135571 | 10,64802 | ,000000 |

$$product = 12.55 + 1.44 fonds$$
$$R^2 = RI = 0.897, \quad s = 2.68$$

Коэффициент детерминации увеличился с 0.597 до 0.897, значение $s$ уменьшилось с 5.01 до 2.68. Действительно, подгонка улучшилась

## 4 Множественная регрессия

Обобщением линейной регрессионной модели с двумя переменными является многомерная регрессионная модель (или модель множественной регрессии). Пусть $n$ раз измерены значения факторов $x_1, x_2, ..., x_k$ и соответствующие значения переменной $y$; предполагается, что (0)

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \varepsilon_i, \quad i \in [1, n] \cap \mathbb{N}$$

(второй индекс у x относится к номеру фактора, а первый - к номеру наблюдения); предполагается также, что (1)

$$M\varepsilon_i = 0, \quad M\varepsilon_i^2 = \sigma^2$$
$$M(\varepsilon_i \varepsilon_j) = 0, \quad i \neq j$$

т.е. $\varepsilon_i$ - некоррелированные случайные величины. Первое соотношения удобно записывать в матричной форме:

$$Y = X\beta + \varepsilon \qquad (13),$$

где $Y = (y_1, ..., y_k)^T$ - вектор-столбец значений зависимой переменной, $T$ - символ транспонирования $\beta = (\beta_0, \beta_1, ..., \beta_k)^T$ - вектор-столбец (размерности $k$) неизвестных коэффициентов регрессии, $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^T$ - вектор случайных отклонений,

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

-матрица $n \times (k+1)$; в $i$ - й строке $(1, x_{i1}, ..., x_{ik})$ находятся значения независимых переменных в $i$-м наблюдении первая переменная - константа, равная 1.

**Оценка коэффициентов регрессии**

Построим оценку для вектора $\widehat{\beta}$ так, чтобы вектор оценок $\widehat{Y} = X\beta$ зависимой переменной минимально (в смысле квадрата нормы разности) отличался от вектора $Y$ заданных значений:

$$||Y - \widehat{Y}||^2 = ||Y - X\beta||^2 \to \min_{\widehat{\beta}}$$

Решением является (если ранг матрицы равен $k+1$) оценка (2)

$$\widehat{\beta} = (X^T X)^{-1} X^T Y$$

Нетрудно проверить, что она несмещенная. Ковариационная (дисперсионная) матрица равна

$$D\widehat{\beta} = (\widehat{\beta} - \beta)(\widehat{\beta} - \beta)^T = \sigma^2 (X^T X)^{-1} = \sigma^2 Z$$

**Теорема Гаусса - Маркова**. В условиях (1) оценка (2) является наилучшей (в смысле минимума дисперсии) оценкой в классе линейных несмещенных оценок.

**Оценка дисперсии $\sigma^2$ ошибок**

Обозначим

$$e = Y - \widehat{Y} = Y - X\widehat{\beta} = \left(I - X(X^T X)^{-1} X^T\right) Y = BY$$

вектор остатков (или невязок); $B = I - X(X^T X)^{-1} X^T$ - матрица; можно проверить, что $B^2 = B$. Для остаточной суммы квадратов $||e||^2$ справедливо соотношение

$$M||e||^2 = M \sum_{i=1}^{n} e_i^2 = (n - k - 1)\sigma^2,$$

откуда следует, что несмещенной оценкой для $\sigma^2$ является

$$s^2 = \frac{||e||^2}{n-k-1} = \frac{Y^T B Y}{n-k-1}$$

Если предположить, что $\varepsilon_i$ нормально распределены, то справедливы следующие свойства оценок:

1. $(n-k-1)\dfrac{s^2}{\sigma^2} \sim \chi^2_{n-k-1}$
2. Оценки $\widehat{\beta}$ и $s^2$ независимы
3. Как и в случае простой регрессии, справедливо соотношение:

$$T_{ss} = R_{ss} + E_{ss},$$

в векторном виде:

$$||Y - \overline{Y}||^2 = ||Y - \widehat{Y}||^2 + ||\widehat{Y} - \overline{Y}||^2$$

Поделив обе части на полную вариацию игреков

$$T_{ss} = \sum_{i=1}^{n} (y_i - \overline{y})^2,$$

получим коэффициент детерминации

$$R^2 = \frac{R_{ss}}{T_{ss}} = 1 - \frac{||Y - \widehat{Y}||^2}{||Y - \overline{Y}||^2}$$

Коэффициент $R^2$ показывает качество подгонки регрессионной модели к наблюдённым значениям $y_i$. Если $R^2 = 0$, то регрессия $Y$ на $x_1, ..., x_k$ не улучшает качество предсказания $y_i$ по сравнению с тривиальным предсказанием. Другой крайний случай $R^2 = 1$ означает точную подгонку: все $\varepsilon_i = 0$, т.е. все точки наблюдений лежат на регрессионной плоскости. Однако, значение $R^2$ возрастает с ростом числа переменных (регрессоров) в регрессии, что не означает улучшения качества предсказания, и потому вводится скорректированный (*adjusted*) коэффициент детерминации

$$R^2_{adj} = 1 - \frac{||Y - \widehat{Y}||^2/(n-k-1)}{||Y - \overline{Y}||/(n-1)} = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

Его использование более корректно для сравнения регрессий при изменении числа переменных (регрессоров).

**Доверительные интервалы для коэффициентов регрессии.** Стандартной ошибкой оценки $\hat{\beta}_j$ является величина $\sigma\sqrt{z_{jj}}$ , оценка для которой

$$s_j = s\sqrt{z_{jj}}\,, \qquad j = 0, 1, ..., k,$$

где $z_{jj}$- диагональный элемент матрицы $Z$. Если ошибки $\varepsilon_i$ распределены нормально, то, в силу свойств 1) и 2), приведенных выше, статистика

$$t = \frac{\left(\hat{\beta}_j - \beta_j\right)/\sigma\sqrt{z_{jj}}}{s/\sigma} = \frac{\hat{\beta}_j - \beta_j}{s_j} \quad \sim \quad t(n-k-1)$$

распределена по закону Стьюдента с ($n$ - $k$ - 1) степенями свободы, и потому неравенство

$$\left|\hat{\beta}_j - \beta_j\right| \le t_p\, s_j\,,$$

где $t_p$ - квантиль уровня $(1 + P_Д) / 2$ этого распределения, задает доверительный интервал для $\beta_j$ с уровнем доверия $P_Д$.

**Проверка гипотезы о нулевых значениях коэффициентов регрессии.** Для проверки гипотезы $H_0$ об отсутствии какой бы то ни было линейной связи между $y$ и совокупностью факторов, $H_0$: $\beta_1 = \beta_2 = ... = \beta_k = 0$, т.е. об одновременном равенстве нулю всех коэффициентов, кроме коэффициента $\beta_0$ при константе, используется статистика

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)} = \frac{R_{ss}}{E_{ss}} \cdot \frac{(n-k-1)}{k} = \frac{\sum_i (\hat{y}_i - \bar{y})^2 \Big/ k}{\sum_i e_i^2 \Big/ (n-k-1)}\,,$$

распределенная, если $H_0$ верна, по закону Фишера с $k$ и $n$ - $k$ - 1 степенями свободы. $H_0$ отклоняется, если

$$F > F_\alpha(k,\, n - k - 1),$$

где $F_\alpha$ - квантиль уровня 1 - $\alpha$.

## Отбор наиболее существенных объясняющих переменных

Различные регрессии (с различным набором переменных) можно сравнивать по скорректированному коэффициенту детерминации $R^2_{adj}$: принять тот вариант регрессии, для которого он максимален

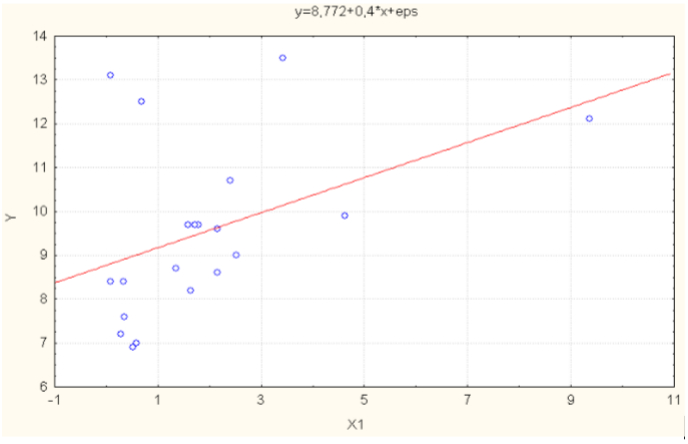
# 5 Пример

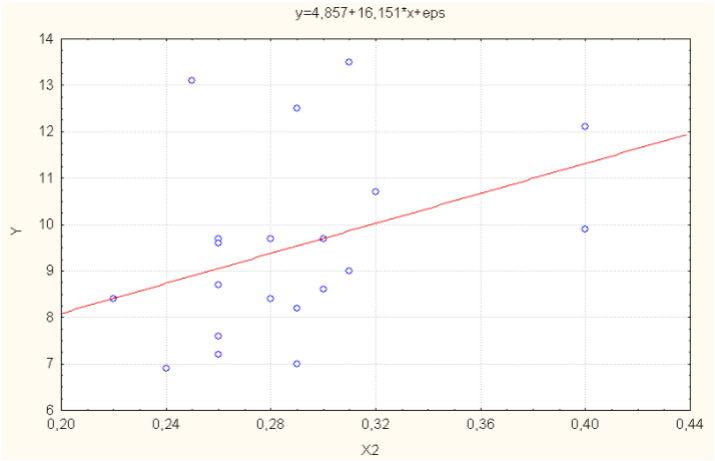
Исследуется зависимость урожайности у зерновых культур ( ц/га ) от ряда факторов (переменных) сельскохозяйственного производства, а именно,
$x_1$ - число тракторов на 100 га;
$x_2$ - число зерноуборочных комбайнов на 100 га;
$x_3$ - число орудий поверхностной обработки почвы на 100 га;
$x_4$ - количество удобрений, расходуемых на гектар (т/га);
$x_5$ - количество на гектар (ц/га) химических средств защиты растений.
Исходные данные для 20 районов области приведены в таблице.

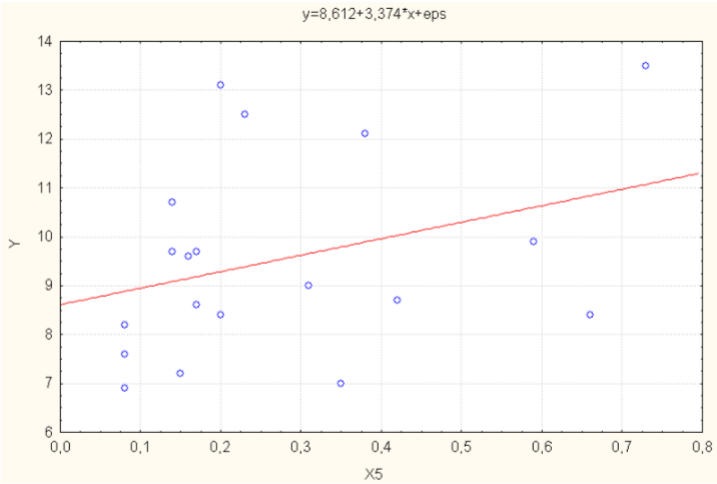
| | $y$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ |
|---|---|---|---|---|---|---|
| 1 | 9.7 | 1.59 | .26 | 2.05 | .32 | .14 |
| 2 | 8.4 | .34 | .28 | .46 | .59 | .66 |
| 3 | 9.0 | 2.53 | .31 | 2.46 | .30 | .31 |
| 4 | 9.9 | 4.63 | .40 | 6.44 | .43 | .59 |
| 5 | 9.6 | 2.16 | .26 | 2.16 | .39 | .16 |
| 6 | 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 |

Здесь мы располагаем выборкой объема $n = 20$; число независимых переменных (факторов) $k = 5$. Матрица должна содержать 6 столбцов размерности 20; первый столбец состоит из единиц, а столбцы со 2-го по 6-й представлены соответственно столбцами 3÷7 таблицы. Специальный анализ (здесь не приводимый) технологии сбора исходных данных показал, что допущения (1) могут быть приняты в качестве рабочей гипотезы, поэтому можем записать уравнения статистической связи между $y_i$ и $Xi = (x_{i1}, x_{i2}, ..., x_{i5}), \quad i \in [1, n] \cap \mathbb{N}$ в виде (0).

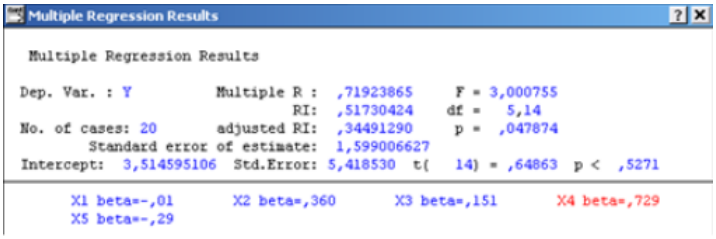
Предварительно визуально оценим имеющиеся данные, построив несколько диаграмм рассеяния:



Иногда такой просмотр позволяет увидеть основную зависимость. В нашем примере этого нет.
Выполним регрессионный анализ.
Для начала посмотрим на зависимость $y$ от всех пяти факторов.




Regression Summary for Dependent Variable: Y

Continue...

R= ,71923865 RI= ,51730424 Adjusted RI= ,34491290
F(5,14)=3,0008 p<,04787 Std.Error of estimate: 1,5990

| N=20 | BETA | St. Err. of BETA | B | St. Err. of B | t(14) | p-level |
|---|---|---|---|---|---|---|
| Intercpt | | | 3,51460 | 5,41853 | ,648625 | ,527078 |
| X1 | -,006596 | 1,002443 | -,00613 | ,93167 | -,006580 | ,994843 |
| X2 | ,359977 | ,498031 | 15,54246 | 21,50311 | ,722800 | ,481704 |
| X3 | ,150640 | 1,141174 | ,10990 | ,83254 | ,132004 | ,896859 |
| X4 | ,728616 | ,251328 | 4,47458 | 1,54345 | 2,899065 | ,011664 |
| X5 | -,288692 | ,304031 | -2,93251 | 3,08833 | -,949546 | ,358448 |

В окне *Mult. Regr. Results* имеем основные результаты: коэффициент детерминации $R^2 = 0.517$; для проверки гипотезы $H_0$ об отсутствии какой бы то ни было линейной связи между переменной $y$ и совокупностью факторов определена статистика $F = 3.00$; это значение соответствует уровню значимости $p = 0.048$ согласно распределению $\mathcal{F}(5, 14)$ с $df = 5$ и 14 степенями свободы. поскольку значение $p$ весьма мало, гипотеза $H_0$ отклоняется.

В заголовке окна *Regression summary* повторены результаты предыдущего окна; в столбце  указаны оценки неизвестных коэффициентов $\widehat{\beta}_j$. Таким образом, оценка $\widehat{f}(x)$ неизвестной функции регрессии $f(x)$ в данном случае:

$$\widehat{f}(x) = 3.51 - 0.06x_1 + 15.5x_2 + 0.11x_3 + 4.47x_4 - 2.93x_5$$

В столбце *St. Err. of B* указаны стандартные ошибки $s_j$ оценок коэффициентов; видно, что стандартные ошибки в оценке всех коэффициентов, кроме $\beta_4$, превышают значения самих коэффициентов, что говорит о статистической ненадежности последних. В столбце $t(14)$ -значение статистики Стьюдента для проверки гипотезы о нулевом значении соответствующих коэффициентов; в столбце *p-level* -уровень значимости отклонения этой гипотезы; достаточно малым (0.01) этот уровень является только для коэффициента при $x_4$ . Только переменная $x_4$ - количество удобрений, подтвердила свое право на включение в модель. В то же время проверка гипотезы об отсутствии какой бы то ни было линейной связи между $y$ и $(x_1, ..., x_5)$ с помощью статистики $F$ (об этом сказано выше)

$$F = 3.00, \quad p = 0.048$$

говорит о том, что следует продолжить изучение линейной связи между $y$ и $(x_1, ..., x_5)$, анализируя как их содержательный смысл, так и матрицу парных корреляций:

Correlations (rehrun2.sta)

| Continue... | X1 | X2 | X3 | X4 | X5 | Y |
|---|---|---|---|---|---|---|
| X1 | 1,000000 | ,854254 | ,977908 | ,110444 | ,341013 | ,430250 |
| X2 | ,854254 | 1,000000 | ,881920 | ,026852 | ,459592 | ,374079 |
| X3 | ,977908 | ,881920 | 1,000000 | ,029819 | ,277923 | ,403153 |
| X4 | ,110444 | ,026852 | ,029819 | 1,000000 | ,570629 | ,577310 |
| X5 | ,341013 | ,459592 | ,277923 | ,570629 | 1,000000 | ,332137 |
| Y | ,430250 | ,374079 | ,403153 | ,577310 | ,332137 | 1,000000 |

Из матрицы видно, что $x_1, x_2, x_3$ (оснащенность техникой) сильно коррелированы (парные коэффициенты корреляции 0.854, 0.882 и 0.978), т.е. имеет место дублирование информации, и потому, по-видимому, есть возможность перехода от исходного числа признаков (переменных) к меньшему.

Сравним различные регрессии, пошагово отбирая переменные.
1-й шаг. Найдем одну наиболее информативную переменную. При $k = 1$ величина $R^2$ совпадает с квадратом обычного (парного) коэффициента корреляции $R^2 = r^2(y, x)$, из матрицы корреляций находим:

$$\max_{1 \le j \le 5} r^2(y, x_i) = r^2(y, x_4) = 0.577^2 = 0.333,$$

так что в классе однофакторных регрессионных моделей наиболее информативным предиктором (предсказателем) является $x_4$ - количество удобрений. Вычислим скорректированный (*adjusted*) коэффициент детерминации: $R^2_{adj}(1) = 0.296$

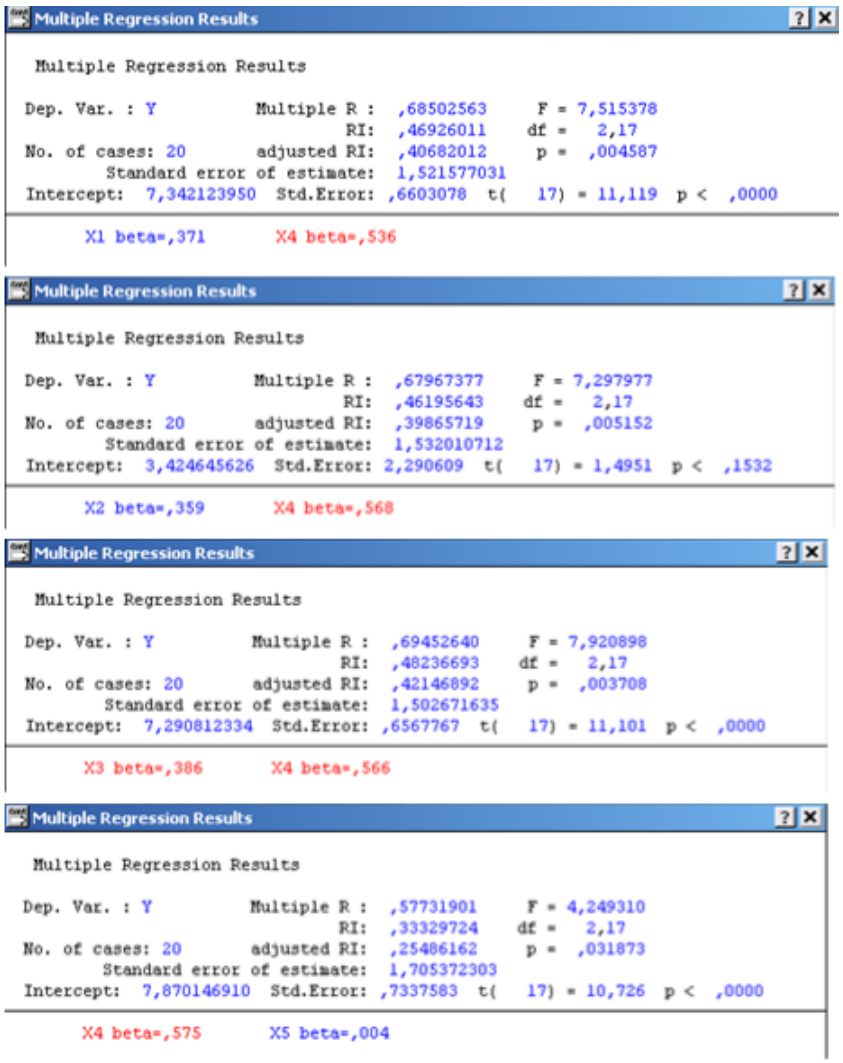
Multiple Regression Results

Multiple Regression Results

Dep. Var. : Y — Multiple R : ,57730960 — F = 8,998098
RI: ,33328637 — df = 1,18
No. of cases: 20 — adjusted RI: ,29624673 — p = ,007691
Standard error of estimate: 1,657337484
Intercept: 7,874628680 Std.Error: ,6633540 t( 18) = 11,871 p < ,0000

X4 beta=,577

2-й шаг. Среди всевозможных пар $(x_4, x_j), \quad j = 1, 2, 3, 5$, выбирается наиболее информативная (в смысле $R^2$ или, что то же самое, в смысле $R^2_{adj}$ ) пара:




$$R^2_{adj}(x_4, x_1) = 0.406 \qquad R^2_{adj}(x_4, x_2) = 0.399$$
$$R^2_{adj}(x_4, x_3) = 0.421 \qquad R^2_{adj}(x_4, x_5) = 0.255$$

откуда видно, что наиболее информативной парой является $(x_4, x_3)$, которая дает

$$R^2_{adj}(2) = \max R^2_{adj}(x_4, x_j) = 0.421$$

Вычислим оценку уравнения регрессии урожайности по факторам $x_3$ и $x_4$.

Regression Summary for Dependent Variable: Y

Continue...

R= ,69452640 RI= ,48236693 Adjusted RI= ,42146892
F(2,17)=7,9209 p<,00371 Std.Error of estimate: 1,5027

| N=20 | BETA | St. Err. of BETA | B | St. Err. of B | t(17) | p-level |
|---|---|---|---|---|---|---|
| Intercpt | | | 7,290812 | ,656777 | 11,10090 | ,000000 |
| X3 | ,386281 | ,174574 | ,281812 | ,127361 | 2,21271 | ,040889 |
| X4 | ,565791 | ,174574 | 3,474635 | 1,072094 | 3,24098 | ,004804 |

$$\widehat{f}(x_3, x_4) = 7.29 + 0.28x_3 + 3.47x_4 \qquad (2)$$

0.66, 0.13, 1.07 - стандартные ошибки, взятые из столбца *Std. Err. of B* таблицы *Regression Results* для варианта независимых переменных $(x_3, x_4)$. Все три коэффициента статистически значимо отличаются от нуля при уровне значимости $\alpha = 0.05$, что видно из столбца *p-level* той же таблицы.

3-й шаг.

Среди всевозможных троек $(x_4, x_3, x_j)$, $j = 1, 2, 5$, выбираем аналогично наиболее информативную:

Multiple Regression Results

Multiple Regression Results

Dep. Var. : Y  Multiple R : ,70585644  F = 5,295776
RI: ,49823331  df = 3,16
No. of cases: 20  adjusted RI: ,40415205  p = ,009975
Standard error of estimate: 1,524995153
Intercept: 7,407576618 Std.Error: ,6864510 t( 16) = 10,791 p < ,0000

X3 beta=,429  X4 beta=,657  X5 beta=-,16

Видим, что наиболее информативной тройкой является $(x_4, x_3, x_5)$, которая дает

$$R^2_{adj}(3) = 0.404$$

что меньше, чем $R^2_{adj}(2) = 0.421$; это означает, что третью переменную в модель включать нецелесообразно, т.к. она не повышает значение (более того, уменьшает). Итак, результатом анализа является (2)

# 6 Нелинейная зависимость

Связь между признаком $x$ и $y$ может быть нелинейной, например, в виде полинома:

$$y = P_k(x) + \varepsilon,$$

где $P_k(x) = \beta_0 + \beta_1 x + ... + \beta_k x^k$, $k$ - степень полинома, $\varepsilon$ - случайная составляющая, $\varepsilon = 0, D\varepsilon = \sigma^2$.
Для имеющихся данных $(x_i, y_i)$, $i = 1, ..., n$, можно записать

$$y_i = \beta_0 + \beta_1 x_i + ... + \beta_k x_i^k + \varepsilon$$

или в матричной форме

$$Y = X\beta + \varepsilon,$$

где

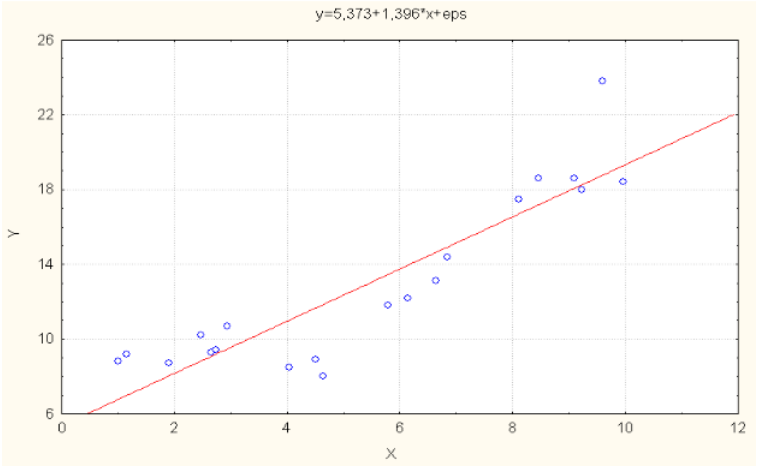
$$X = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^k \\ 1 & x_2 & x_2^2 & \cdots & x_2^k \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^k \end{pmatrix}$$

Имеем задачу (13), и потому все формулы п.2. оказываются справедливыми и в этом случае. Слово “линейный” в названии “линейный регрессионный анализ” означает линейность относительно параметров $\beta_j$ , но не относительно факторов $x_j$.

Пример. Имеются эмпирические данные о зависимости $y$ - выработки на одного работника доменного производства от $x$ - температуры дутья; данные приведены в таблице в условных единицах.

| № | X | Y | № | X | Y |
|---|---|---|---|---|---|
| 1 | 1.01 | 8.8 | 11 | 5.80 | 11.8 |
| 2 | 1.15 | 9.2 | 12 | 6.14 | 12.2 |
| 3 | 1.91 | 8.7 | 13 | 6.64 | 13.1 |
| 4 | 2.47 | 10.2 | 14 | 6.85 | 14.4 |
| 5 | 2.66 | 9.3 | 15 | 8.11 | 17.5 |
| 6 | 2.74 | 9.4 | 16 | 8.47 | 18.6 |
| 7 | 2.93 | 10.7 | 17 | 9.09 | 18.6 |
| 8 | 4.04 | 8.5 | 18 | 9.23 | 18.0 |
| 9 | 4.50 | 8.9 | 19 | 9.59 | 23.8 |
| 10 | 4.64 | 8.0 | 20 | 9.96 | 18.4 |

Образуем таблицу $4\nu \times 20c$, в первые 2 столбца поместим исходные данные $x$ и $y$. В третьем столбце поместим значения нового фактора $x2$ квадратов температур, в четвертом - $x3$ третьих степеней температур. Сначала оценим имеющиеся данные визуально, с помощью диаграммы рассеяния.




Видим, что зависимость, возможно, нелинейная. Построим несколько регрессий: 1) первой степени: $y = \beta_0 + \beta_1 x$; получим:

Regression Summary for Dependent Variable: Y

Continue...

R= ,89808430 RI= ,80655541 Adjusted RI= ,79580849
F(1,18)=75,050 p<,00000 Std.Error of estimate: 2,0992

| N=20 | BETA | St. Err. of BETA | B | St. Err. of B | t(18) | p-level |
|---|---|---|---|---|---|---|
| Intercpt | | | 5,372930 | ,988060 | 5,437857 | ,000036 |
| X | ,898084 | ,103667 | 1,395732 | ,161112 | 8,663135 | ,000000 |

$$y = 5.37 + 1.40x,\ \textit{Std. Err. of } B\text{: } 0.98, 0.16$$
$$R^2_{adj} = 0.796, s = 2.09$$

2) второй степени: $y = \beta_0 + \beta_1 x + \beta_2 x^2$; получим:

Regression Summary for Dependent Variable: Y

Continue...

R= ,94994734 RI= ,90239994 Adjusted RI= ,89091758
F(2,17)=78,590 p<,00000 Std.Error of estimate: 1,5343

| N=20 | BETA | St. Err. of BETA | B | St. Err. of B | t(17) | p-level |
|---|---|---|---|---|---|---|
| Intercpt | | | 9,956839 | 1,334237 | 7,46257 | ,000001 |
| X | -,581944 | ,370072 | -,904412 | ,575137 | -1,57252 | ,134256 |
| X2 | 1,512061 | ,370072 | ,208153 | ,050945 | 4,08586 | ,000770 |

$$y = 9.96 - 0.9x + 0.21x^2,\ \textit{Std. Err. of } B\text{: } 1.33, 0.57, 0.05$$
$$R^2_{adj} = 0.891, s = 1.53$$

Эта регрессия лучше предыдущей в смысле $R^2_{adj}$ и $s$. Однако, возможно, регрессия третьей степени окажется лучше?

3) третьей степени: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$; получим:

Regression Summary for Dependent Variable: Y

Continue...

R= ,95220628 RI= ,90669680 Adjusted RI= ,88920245
F(3,16)=51,828 p<,00000 Std.Error of estimate: 1,5463

| N=20 | BETA | St. Err. of BETA | B | St. Err. of B | t(16) | p-level |
|---|---|---|---|---|---|---|
| Intercpt | | | 11,60650 | 2,345520 | 4,94837 | ,000145 |
| X | -1,49268 | 1,124621 | -2,31981 | 1,747798 | -1,32727 | ,203048 |
| X2 | 3,75005 | 2,633718 | ,51624 | ,362563 | 1,42386 | ,173696 |
| X3 | -1,36097 | 1,585487 | -,01887 | ,021978 | -,85839 | ,403351 |

$$y = 11.6 - 2.32x + 0.52x^2 - 0.02x^3,\ \textit{Std. Err. of } B\text{: } 2.35, 1.75, 0.36, 0.02$$
$$R^2_{adj} = 0.889, s = 1.55$$

Поскольку степень увеличилась без увеличения $R^2_{adj}$, от регрессии третьей степени отказываемся в пользу второй степени. Однако гипотеза о нулевом значении $\beta_1$ в регрессии второй степени не отклоняется ($p\text{-}level = 0.1$), и потому построим

4) регрессию $y = \beta_0 + \beta_2 x^2$ без линейного члена; получим:

Regression Summary for Dependent Variable: Y

Continue...

R= ,94244528 RI= ,88820311 Adjusted RI= ,88199217
F(1,18)=143,01 p<,00000 Std.Error of estimate: 1,5959

| N=20 | BETA | St. Err. of BETA | B | St. Err. of B | t(18) | p-level |
|---|---|---|---|---|---|---|
| Intercpt | | | 8,025412 | ,542069 | 14,80516 | ,000000 |
| X2 | ,942445 | ,078810 | ,129739 | ,010849 | 11,95852 | ,000000 |

$$y = 8.03 + 0.13x^2,\ \textit{Std. Err. of } B\text{: } 0.54, 0.01$$
$$R^2_{adj} = 0.882, s = 1.6$$

Сравнивая ее по $R^2_{adj}$ и $s$ с регрессией для второй степени, отдаем предпочтение регрессии для второй степени, поскольку ошибка $s$ прогноза меньше.

# 7 Нелинейная зависимость – обобщение

Предполагается, что связь между факторами $(x_1, ..., x_p)$ и $y$ выражается следующим образом:

$$y = \beta_0 + \beta_1\varphi_1(x_1, ..., x_p) + ... + \beta_k\varphi_k(x_1, ..., x_p) + \varepsilon,$$

где $\varphi_j(\cdot, ..., \cdot)$ - система некоторых функций. Имеется $n$ наблюдений при различных значениях $x \equiv ((x_1, ..., x_p)$: $x^1, ..., x^n$ ; тогда

$$y = \beta_0 + \sum_{j=1}^{k} \beta_j\varphi_j(x^i) + \varepsilon_i,\ i = 1, ..., n,$$

или в матричной форме

$$Y = X\beta + \varepsilon,$$

где - матрица $n \times (k + 1)$, в $i$-й строке которой $(1, \varphi_1(x^i), \varphi_2(x^i), ..., \varphi_k(x^i))$; $y$, $\beta$ , $\varepsilon$, как в (13). Получили задачу (13), и потому все формулы п.2 оказываются справедливыми.

Пример. Имеется 20 наблюдений по некоторому технологическому процессу химического производства; $x$, $y$ - изменяемое содержание двух веществ , $z$ - контролируемый параметр получаемого продукта. Полагая, что

$$z = P(x, y) + \varepsilon,$$

где $P(x, y) = \beta_0 + \beta_1 x + \beta_2 y + \beta_3 x^2 + \beta_4 xy + \beta_5 y^2$ - многочлен второй степени, $\varepsilon$ - случайная составляющая, $\varepsilon = 0, D\varepsilon = \sigma^2$, необходимо оценить функцию $P(x, y)$ и найти точку ее минимума. Исходные данные приведены в таблице.

| $i$ | $x_i$ | $y_i$ | 7 $z_i$ |
|---|---|---|---|
| 1 | -3 | -2 | 160 |
| 2 | -3 | 1 | 61.4 |
| 3 | -3 | 3 | 0.5 |
| 4 | -2 | -3 | 148.8 |
| 5 | -2 | 0 | 86.5 |
| 6 | -2 | 2 | 45 |
| 7 | -1 | -2 | 121.2 |
| 8 | -1 | 3 | 18 |
| 9 | 0 | -3 | 74.2 |
| 10 | 0 | -1 | 110.2 |
| 11 | 0 | 2 | 99.6 |
| 12 | 1 | -1 | 107.9 |
| 13 | 1 | 1 | 94.5 |
| 14 | 1 | 3 | 115.4 |
| 15 | 2 | -3 | 17.1 |
| 16 | 2 | 1 | 105.4 |
| 17 | 2 | -3 | 86.9 |
| 18 | 3 | -2 | 7.7 |
| 19 | 3 | 0 | 60.9 |
| 20 | 3 | 2 | 112.2 |

Образуем таблицу $6\nu \times 20c$, в 3 столбца которой ввести исходные данные. Образуем новые факторы - столбцы, соответствующие $x^2$, $xy$, $y^2$, и вычислим их значения.

Проведем регрессионный анализ:

Regression Summary for Dependent Variable: Z

Continue...

R= ,91126516 RI= ,83040420 Adjusted RI= ,76983427
F(5,14)=13,710 p<,00006 Std.Error of estimate: 21,848

| N=20 | BETA | St. Err. of BETA | B | St. Err. of B | t(14) | p-level |
|---|---|---|---|---|---|---|
| Intercpt | | | 100,1670 | 10,71454 | 9,34869 | ,000000 |
| X | -,126634 | ,111895 | -2,7600 | 2,43873 | -1,13172 | ,276765 |
| Y | -,209164 | ,112401 | -4,3334 | 2,32867 | -1,86088 | ,083892 |
| X2 | -,148392 | ,112427 | -1,8913 | 1,43294 | -1,31990 | ,208043 |
| XY | ,819221 | ,114725 | 8,7830 | 1,22998 | 7,14073 | ,000005 |
| Y2 | -,097215 | ,116105 | -1,2334 | 1,47301 | -,83730 | ,416490 |

Получили регрессию

$$z(x,y) = 100.167 - 2.76x - 4.33y - 1.89x^2 + 8.78xy - 1.23y^2$$

*Std. Err. of B*: $10.71, 2.43, 2.32, 1.43, 1.23, 1.47$

Трехмерный график полученной функции: