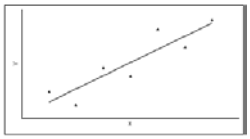
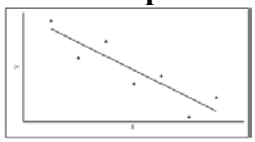
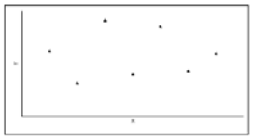
Тема 4 «Проверка гипотезы о значимости уравнения регрессии.»

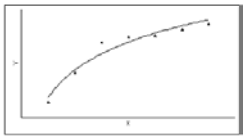
Основные понятия и формулы

Форма связи


линейная положительная

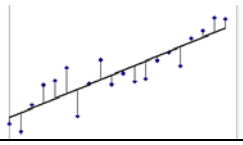

линейная отрицательная


отсутствует

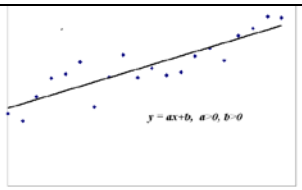

нелинейная

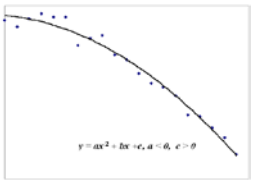
Метод наименьших квадратов (МНК)

$$S = \sum_{i=1}^{n} \left(y_{i\,эксп} - y_{i\,теор}\right)^2 = \sum_{i=1}^{n} \left(y_{i\,эксп} - \varphi(x_i, b_0, b_1, \dots b_p)\right)^2 \to min$$



Регрессионный анализ


линейная регрессия


нелинейная регрессия

$$Y = b_0 + b_1X_1 + \dots + b_kX_k$$

$X_1, \dots, X_k$ – *факторы*

$b_0, \dots, b_k$ - *коэффициенты*

множественная регрессия

В зависимости от типа выбранного уравнения различают *линейную* и *нелинейную* регрессию (в последнем случае возможно дальнейшее уточнение: квадратичная, экспоненциальная, логарифмическая и т.д.). В зависимости от числа взаимосвязанных признаков различают *парную* и *множественную* регрессию. Если исследуется связь между двумя признаками (результативным и факторным), то регрессия называется парной, если между тремя и более признаками – множественной (многофакторной) регрессией.

Парная линейная регрессионная модель.

Уравнение регрессии будет иметь вид: $y_x = ax + b$, где a – *коэффициент регрессии* (показатель наклона линии линейной регрессии).

С помощью метода наименьших квадратов получают формулы, по которым можно вычислять параметры линейной регрессии:

Свободный член b	Коэффициент регрессии a	Коэффициент детерминации
$b = \frac{\overline{y} \cdot \overline{x^2} - \overline{x} \cdot \overline{xy}}{\overline{x^2} - (\overline{x})^2}$	$a = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - (\overline{x})^2}$	$R^2 = \frac{\Sigma(y_i^p - \overline{y})^2}{\Sigma(y_i - \overline{y})^2}$
Проверка гипотезы о значимости уравнения регрессии		
H_0: $R^2 = 0$	H_1: $R^2 > 0$	$F_{набл} = \frac{R^2}{1 - R^2} \cdot \frac{n - p - 1}{p}$
$F_{кр}(\alpha; k_1; k_2)$, $k_1 = p$, $k_2 = n - p - 1$, *(для линейной регрессии p = 1)*		

Для анализа общего качества уравнения регрессии используют коэффициент детерминации R^2, называемый также квадратом коэффициента множественной корреляции. Коэффициент детерминации (мера определенности) всегда находится в пределах интервала [0;1]. Если значение R^2 близко к единице, это означает, что построенная модель объясняет почти всю изменчивость соответствующих переменных. И наоборот, значение R^2 близкое к нулю, означает плохое качество построенной модели.

При высоком значении коэффициента детерминации ($R^2 \geq 75\%$) можно делать прогноз $y^* = f(x^*)$ для конкретного значения x^* в пределах диапазона исходных данных. При прогнозах значений, не входящих в диапазон исходных данных, справедливость

полученной модели гарантировать нельзя. Это объясняется тем, что может проявиться влияние новых факторов, которые модель не учитывает.

Оценка значимости уравнения регрессии осуществляется с помощью критерия Фишера. При условии справедливости нулевой гипотезы критерий имеет распределение Фишера с числом степеней свободы $k_1 = p$, $k_2 = n - p - 1$ (для парной линейной регрессии $p = 1$). Если нулевая гипотеза отклоняется, то уравнение регрессии считается статистически значимым. Если нулевая гипотеза не отклоняется, то признается статистическая незначимость или ненадежность уравнения регрессии.

Если $F_{набл} > F_{кр}$, то нулевая гипотеза отвергается. ***Уравнение регрессии (линейной модели) статистически значимо.***

<u>Пример 1.</u> В механическом цехе анализируется структура себестоимости продукции и доля покупных комплектующих. Было отмечено, что стоимость комплектующих зависит от времени их поставки. В качестве наиболее важного фактора, влияющего на время поставки, выбрано пройденное расстояние. Провести регрессионный анализ данных о поставках:

Расстояние, миль	*3,5*	*2,4*	*4,9*	*4,2*	*3,0*	*1,3*	*1,0*	*3,0*	*1,5*	*4,1*
Время, мин	*16*	*13*	*19*	*18*	*12*	*11*	*8*	*14*	*9*	*16*

Для проведения регрессионного анализа:

1. *построить график исходных данных, приближенно определить характер зависимости;*
2. *выбрать вид функции регрессии и определить численные коэффициенты модели методом наименьших квадратов и направление связи;*
3. *оценить силу регрессионной зависимости с помощью коэффициента детерминации;*
4. *оценить значимость уравнения регрессии;*
5. *сделать прогноз (или вывод о невозможности прогнозирования) по принятой модели для расстояния 2 мили*

Решение

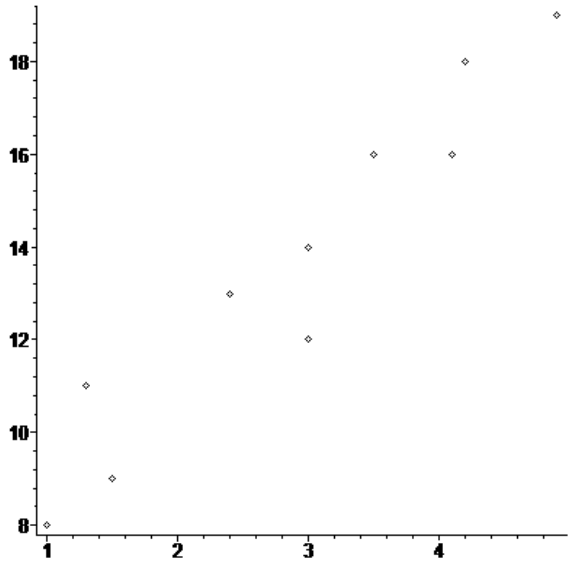
```
> restart;
> with(plots):
> with(CurveFitting):
> X:=[3.5,2.4,4.9,4.2,3.0,1.3,1,3,1.5,4.1];
```

$$X := [3.5, 2.4, 4.9, 4.2, 3.0, 1.3, 1, 3, 1.5, 4.1]$$

```
> Y:=[16,13,19,18,12,11,8,14,9,16];
```

$$Y := [16, 13, 19, 18, 12, 11, 8, 14, 9, 16]$$

```
> q0:=plot([[X[i],Y[i]]$i=1..10],style=point,color=black):
> display(q0);
```



Точки собраны вдоль прямой линии, поэтому можно предположить линейную положительную связь между параметрами.

```
> Digits := 5;
> p1:=LeastSquares(X, Y, x, curve=a*x+b);
```

$$p1 := 5.9135 + 2.6597\,x$$

```
> f1 := unapply(p1,x);
```

$$f1 := x \rightarrow 5.9135 + 2.6597\,x$$

```
> b:=f1(0);
```

$$b := 5.9135$$

```
> a:=f1(1)-b;
```

$$a := 2.6597$$

Вычислим $\bar{y} = \dfrac{\sum n_i y_i}{n}$ и расчетное y_i^P.

```
> N:=10;  SUMY:=0;
```

$$N := 10$$

$$SUMY := 0$$

```
> for i from 1 to N do
 SUMY:=SUMY+Y[i]: end do:
```

Среднее $\bar{y} = \dfrac{\sum n_i y_i}{n}$

```
> MY:=evalf(SUMY/N);
```

$$MY := 13.600$$

Вычислим R^2

```
> RR1:=0; RR2:=0;
```

$$RR1 := 0$$

$$RR2 := 0$$

```
> for i from 1 to N do
d:=X[i];
CALCY[i]:=f1(d);
RR1:=RR1+(CALCY[i]-MY)^2;
RR2:=RR2+(Y[i]-MY)^2;
end do;
```

$$RR1 := 112.40$$

$$RR2 := 122.40$$

```
> RSQ:=RR1/RR2;
```

$$RSQ := 0.91830$$

Таким образом коэффициент детерминации: R^2*=0,92 или 92%. Таким образом, линейная модель объясняет 92% вариации =времени поставки, что означает правильность выбора фактора (расстояния). Не объясняется 8% вариации времени, которые обусловлены остальными факторами, влияющими на время поставки, но не включенными в линейную модель регрессии.*

Вычислим $F_{\text{набл}} = \dfrac{R^2}{1-R^2} \cdot \dfrac{n-p-1}{p}$

```
> pp:=1;
```

$pp := 1$

```
> FN:=RSQ*(N-pp-1)/((1-RSQ)*pp);
```

$FN := 46.873$

По таблице находим $F_{кр}(\alpha; k_1; k_2)$ $\alpha = 0{,}05$, $k_1 = p, k_2 = n - p - 1$, *(для линейной регрессии p = 1)*

Т.к. $F_{набл} > F_{кр}(0{,}05;1;10-1-1) = 5{,}32$ *– уравнение регрессии (линейной модели) статистически значимо.*

5. Решим задачу прогнозирования. Поскольку коэффициент детерминации R^2 *имеет достаточно высокое значение и расстояние 2 мили, для которого надо сделать прогноз, находится в пределах диапазона исходных данных, то можно сделать прогноз:*

```
> y:=f1(2);
```

$y := 11.233$

Нелинейная регрессия

Пример 2. Для массива экспериментальных данных построить возможные уравнения нелинейной регрессии и по максимальному коэффициенту детерминации найти наилучшее уравнение нелинейной регрессии.

X	Y
1	10
2	13,4
3	15,4
4	16,5
5	18,6
6	19,1

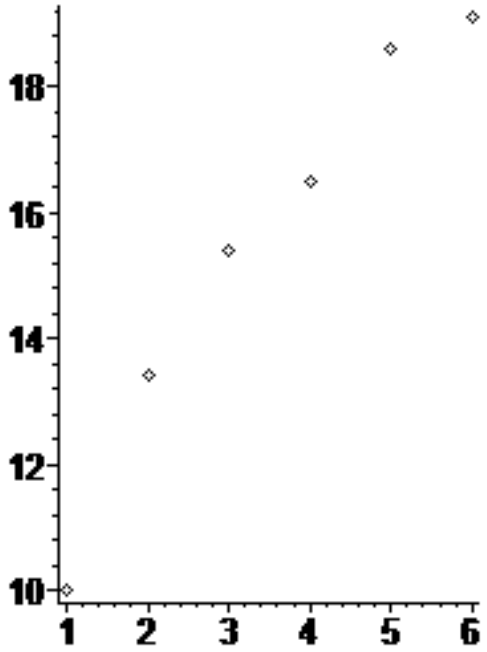
Решение

```
> restart;
> with(plots):
> with(CurveFitting):
> X:=[1,2,3,4,5,6];
> Y:=[10,13.4,15.4,16.5,18.6,19.1];
```

$X := [1, 2, 3, 4, 5, 6]$

$Y := [10, 13.4, 15.4, 16.5, 18.6, 19.1]$

```
> q0:=plot([[X[i],Y[i]]$i=1..6],style=point,color=black):
> display(q0);
```




```
> N:=6;  SUMY:=0;
```

$$N := 6$$

$$SUMY := 0$$

```
> for i from 1 to N do
 SUMY:=SUMY+Y[i]: end do:
> SUMY;
```

$$93.0$$

```
> MY:=evalf(SUMY/N);
```

$$MY := 15.50000000$$

```
> Digits := 5;
```

$$Digits := 5$$

```
> p1:=LeastSquares(X, Y, x, curve=a*x+b);
```

$$p1 := 9.2800 + 1.7771\,x$$

```
> f1 := unapply(p1,x);
```

$$f1 := x \rightarrow 9.2800 + 1.7771\,x$$

```
> p2:=LeastSquares(X, Y, x, curve=a*x^2+b*x+c);
```

$$p2 := 6.9300 + 3.5396\,x - 0.25179\,x^2$$

```
> f2:= unapply(p2,x);
```

$$f2 := x \rightarrow 6.9300 + 3.5396\,x - 0.25179\,x^2$$

```
>
> p3:=LeastSquares(X, Y, x, curve=a+b*ln(x));
```

$$p3 := 9.8753 + 5.1296\,\ln(x)$$

```
> f3 := unapply(p3,x);
```

$$f3 := x \rightarrow 9.8753 + 5.1296\,\ln(x)$$

нелинейная апроксимация

```
> with(Statistics):
> X := Vector([1, 2, 3, 4, 5, 6], datatype=float):
Y := Vector([10,13.4,15.4,16.5,18.6,19.1], datatype=float):
> Digits := 5;
```

$$Digits := 5$$

Для сравнения

```
> p2N:=Fit(a+b*t+c*t^2, X, Y, t);
```

$$p2N := 6.9300000000000860 + 3.53964285714285420\ t - 0.251785714285713891\ t^2$$

```
> p4:=Fit(a*exp(b*t), X, Y, t);
```

$$p4 := 10.3639086397654286\ \mathbf{e}^{(0.110198243772794038\ t)}$$

```
> f4:= unapply(p4,t);
```

$$f4 := t \to 10.3639086397654286\ \mathbf{e}^{(0.110198243772794038\ t)}$$

```
> p5:=Fit(a*t^b, X, Y, t);
```

$$p5 := 10.2829969941739883\ t^{0.354496026863205471}$$

```
> f5 := unapply(p5,t);
```

$$f5 := t \to 10.2829969941739883\ t^{0.354496026863205471}$$

```
> RR11:=0; RR2:=0;RR12:=0;RR13:=0;RR14:=0;RR15:=0;
> for i from 1 to N do
d:=X[i];
CALCY1[i]:=f1(d);
CALCY2[i]:=f2(d);
CALCY3[i]:=f3(d);
CALCY4[i]:=f4(d);
CALCY5[i]:=f5(d);
RR11:=RR11+(CALCY1[i]-MY)^2;
RR12:=RR12+(CALCY2[i]-MY)^2;
RR13:=RR13+(CALCY3[i]-MY)^2;
RR14:=RR14+(CALCY4[i]-MY)^2;
RR15:=RR15+(CALCY5[i]-MY)^2;
RR2:=RR2+(Y[i]-MY)^2;
end do;
>RSQ1:=RR11/RR2;RSQ2:=RR12/RR2;RSQ3:=RR13/RR2;RSQ4:=RR14/RR2;RSQ5
:=RR15/RR2;
```

$$RSQ1 := 0.94909$$

$$RSQ2 := 0.98958$$

$$RSQ3 := 0.99183$$

$$RSQ4 := 0.87079$$

$$RSQ5 := 0.98036$$

Вывод: в качестве наилучшего уравнения регрессии выбираем степенную функцию

$$f3 := x \to 9.8753 + 5.1296 \ln(x)$$

Замечание

Функция fit может обеспечивать регрессию и для функций нескольких переменных. В качестве примера приведен пример для функции двух переменных

```
> restart;
> with(stats):
>
p:=fit[leastsquare[[x,y,z],z=a+b*x^2+c*y,{a,b,c}]]\([[1,2,3,5,5]
,[2,4,6,8,8],
[3,5,7,10,11]]);
```

$$p := z = \frac{107}{77} + \frac{37\,x^2}{308} + \frac{235\,y}{308}$$

```
> f:= unapply(p,x,y);
```

$$f := (x, y) \to z = \frac{107}{77} + \frac{37}{308}x^2 + \frac{235}{308}y$$

Задачи для самостоятельной работы

Задача 1. Экспериментальные данные по обкатыванию поверхности шаровым инструментом и шероховатости обработанной поверхности приведены в таблице:

X - сила прижима, *кгс*	50	75	100	125	150
Y – шероховатость, *мкм*	0,60	0,54	0,47	0,40	0,31

Для проведения регрессионного анализа:

1. построить график исходных данных, приближенно определить характер зависимости;
2. выбрать вид функции регрессии и определить численные коэффициенты модели методом наименьших квадратов;
3. оценить силу регрессионной зависимости с помощью коэффициента детерминации;
4. оценить значимость уравнения регрессии;
5. сделать прогноз (или вывод о невозможности прогнозирования) по принятой модели для силы прижима 135 *кгс*.

Задача 2. Для массива экспериментальных данных построить возможные уравнения регрессии и по максимальному коэффициенту детерминации найти наилучшее уравнение регрессии.

X	3	8	5	10	7	6	4	9	1	2
Y	6	5	9	1	8	9	8	4	2	4