Министерство образования и науки Российской Федерации

Федеральное государственное бюджетное

образовательное учреждение высшего образования

ПСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Институт инженерных наук

Кафедра информационно-коммуникационных технологий

ЛАБОРАТОРНАЯ РАБОТА №6

ЛИНЕЙНАЯ РЕГРЕССИЯ

Вариант 13

по дисциплине «Моделирование»

Выполнила: Разгонова Е.В.

Группа: 0432-04

Проверил: Миронов Т.С.

Псков
2021

Задание 6.1. Линейная регрессия

<u>Задание</u>: для заданной в условии выборки вычислите регрессию и найдите доверительные интервалы коэффициентов регрессии и дисперсии для заданной доверительной вероятности. Вычислите коридор и доверительную область регрессии. Изобразите выборку графически на одном графике с линией регрессии. Изобразите графически коридор и доверительную область регрессии.

x	-0.9	-0.8	-0.7	-0.6	-0.5	-0.4	-0.3	-0.2
y	5.892	5.103	5.624	5.197	4.749	4.653	4.253	4.249
x	-0.1	0	0.1	0.2	0.3	0.4	0.5	
y	4.555	3.955	4.076	3.869	3.241	2.782	2.667	

$ORIGIN := 1 \qquad n := 15$

$x := (-0.9\ \ -0.8\ \ -0.7\ \ -0.6\ \ -0.5\ \ -0.4\ \ -0.3\ \ -0.2\ \ -0.1\ \ 0\ \ 0.1\ \ 0.2\ \ 0.3\ \ 0.4\ \ 0.5)$

$y := (5.892\ \ 5.103\ \ 5.624\ \ 5.197\ \ 4.749\ \ 4.653\ \ 4.253\ \ 4.249\ \ 4.555\ \ 3.955\ \ 4.076\ \ 3.869\ \ 3.241\ \ 2.782\ \ 2.667)$

$x := x^T \quad y := y^T$

Точечные оценки мат.ожидания

$Xmean := mean(x) \qquad Xmean = -0.2 \qquad Ymean := mean(y) \qquad Ymean = 4.324$

Точечная несмещённая оценка дисперсии и линейная модель регрессии

$a0 := intercept(x, y) \qquad a0 = 3.918$

$a1 := slope(x, y) \qquad a1 = -2.03$

$i := 1..n$

$lmr := a0 + a1 \cdot x \qquad Dx := \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(y_i - a0 + a1 \cdot x_i\right)^2 \qquad Dx = 4.077$

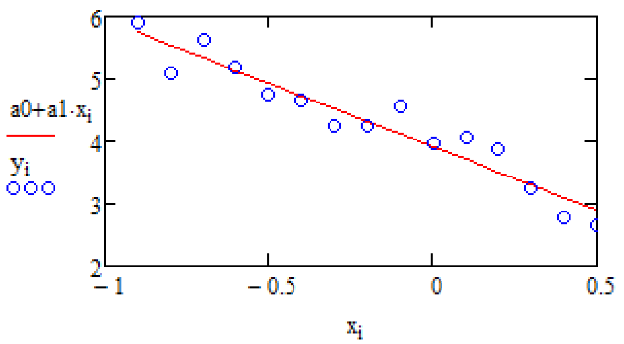

Рисунок 1. Фрагмент рабочего документа MathCAD (1)

Вычисление доверительного интервала для параметра a0

$p := 0.9 \quad \alpha := 1 - p = 0.1$

$t := qt\left(1 - \frac{\alpha}{2}, n - 2\right) \quad t = 1.771 \quad s2 := \frac{1}{n-2} \cdot \sum_{i=1}^{n} \left(y_i - lmr_i\right)^2$

$$a0left := a0 - t \cdot \sqrt{s2} \cdot \sqrt{\frac{1}{n} + \frac{Xmean^2}{\sum_{i=1}^{n} \left(x_i - Xmean\right)^2}} \qquad a0left = 3.776$$

$$a0right := a0 + t \cdot \sqrt{s2} \cdot \sqrt{\frac{1}{n} + \frac{Xmean^2}{\sum_{i=1}^{n} \left(x_i - Xmean\right)^2}} \qquad a0right = 4.0603$$

Рисунок 2. Фрагмент рабочего документа MathCAD (2)

Вычисление доверительного интервала для параметра a1

$$a1left := a1 - \frac{t \cdot \sqrt{s2}}{\sqrt{\sum_{i=1}^{n} \left(x_i - Xmean\right)^2}} \qquad a1left = -2.328$$

$$a1right := a1 + \frac{t \cdot \sqrt{s2}}{\sqrt{\sum_{i=1}^{n} \left(x_i - Xmean\right)^2}} \qquad a1right = -1.732$$

Рисунок 3. Фрагмент рабочего документа MathCAD (3)

Вычисление доверительного интервала для дисперсии

$xleft := qchisq\left(\frac{\alpha}{2}, n - 2\right) \quad xleft = 5.892 \qquad \sigma right := \frac{(n-2) \cdot s2}{xleft} \quad \sigma right = 0.175$

$xright := qchisq\left(1 - \frac{\alpha}{2}, n - 2\right) \quad xright = 22.362 \qquad \sigma left := \frac{(n-2) \cdot s2}{xright} \quad \sigma left = 0.046$

Рисунок 4. Фрагмент рабочего документа MathCAD (4)

Построение доверительного коридора и доверительной области

$$yleft_i := \left(a0 + a1 \cdot x_i\right) - t \cdot \sqrt{s2} \cdot \sqrt{\frac{1}{n} + \frac{\left(x_i - Xmean\right)^2}{\sqrt{\sum_{i=1}^{n} \left(x_i - Xmean\right)^2}}}$$

$$yright_i := \left(a0 + a1 \cdot x_i\right) + t \cdot \sqrt{s2} \cdot \sqrt{\frac{1}{n} + \frac{\left(x_i - Xmean\right)^2}{\sqrt{\sum_{i=1}^{n} \left(x_i - Xmean\right)^2}}}$$

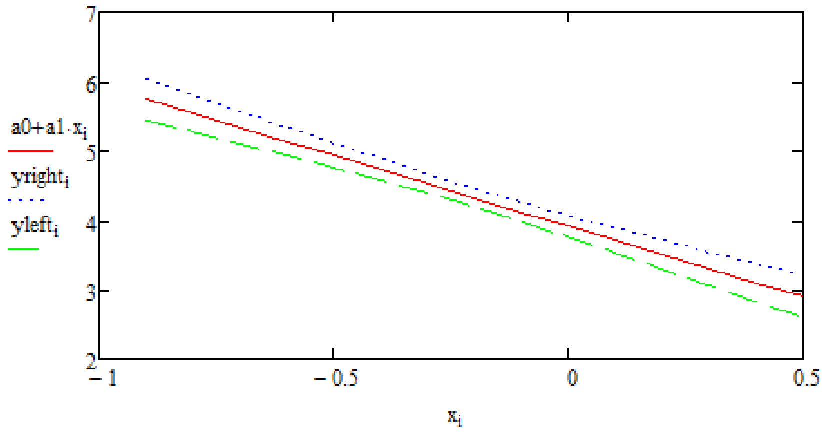

Рисунок 5. Фрагмент рабочего документа MathCAD (5)

$$f := qF(1 - \alpha, 2, n - 2)$$

$$yleft_i := \left(a0 + a1 \cdot x_i\right) - 2 \cdot f \cdot \sqrt{s2} \cdot \sqrt{\frac{1}{n} + \frac{\left(x_i - Xmean\right)^2}{\sqrt{\sum_{i=1}^{n} \left(x_i - Xmean\right)^2}}}$$

$$yright_i := \left(a0 + a1 \cdot x_i\right) + 2 \cdot f \cdot \sqrt{s2} \cdot \sqrt{\frac{1}{n} + \frac{\left(x_i - Xmean\right)^2}{\sqrt{\sum_{i=1}^{n} \left(x_i - Xmean\right)^2}}}$$

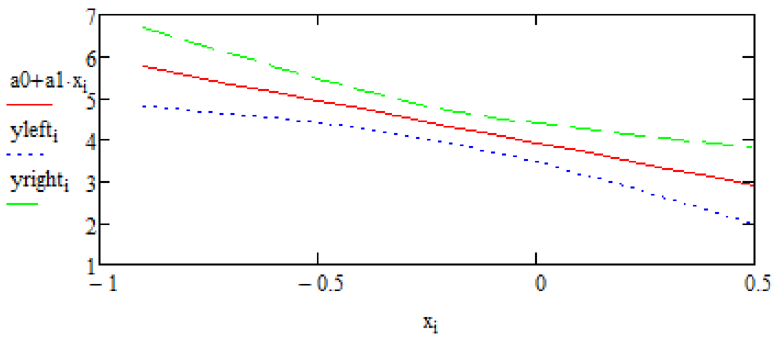

Рисунок 6. Фрагмент рабочего документа MathCAD (6)

Пояснение:

То, как мы определяем и оцениваем параметры выборок задаётся линейной регрессией. В основном она используется для моделирования технологических процессов, где основную роль играет связь между входными и выходными данными. Пусть требуется исследовать зависимость $y(x)$, причем величины y и x измеряются в одних и тех же экспериментах. можно считать, что величина x измеряется точно, в то время как измерение величины y содержит случайные погрешности. Это означает, что погрешность измерения величины x весьма мала по сравнению с погрешностью измерения величины y. Таким образом, результаты эксперимента можно рассматривать как выборочные значения случайной величины $\eta(x)$, зависящей от x, как от параметра.

Регрессией называют зависимость $y(x)$ условного математического ожидания величины $\eta(x)$ от переменной x, т.е. $y(x) = M(^{\eta}/_{x})$. Задача регрессивного анализа состоит в восстановлении функциональной зависимости $y(x)$ по результатам измерений (x_i, y_i), $i = 1,2 \dots, n.$ Эти результаты измерений можно представить в виде линейной функции $y_i = f(x, a_0, a_1, \dots, a_k) + \xi_i$, где $a_0, a_1, \dots, a_k$ – неизвестные коэффициенты регрессии, а величина ξ_i – случайные величины (обычно нормально распределённые, т.е. $M[\xi_i] = 0$ и $D[\xi_i] = \sigma^2$), характеризующие погрешности эксперимента.

Параметры $a_0, a_1, \dots, a_k$ следует выбирать таким образом, чтобы отклонение значений предложенной функции от результатов эксперимента было минимальным. Сами коэффициенты функции определяют методом наименьших квадратов.

$$\sum_{i=1}^{n} {e_i}^2 = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2 \to min$$

Здесь, величина e_i называется невязкой, которая характеризует отклонение фактических значений y_i от теоретических, полученных по уравнению регрессии. Эта величина возникает в результате ошибки эксперимента. Чтобы сумма была минимальной, частные производные коэффициентов должны быть равны 0. Таким образом мы получаем оценку погрешности, а затем, через дифференцирование, получаем точные значения оценок, при которых будут соблюдены указанное выше условие.

Выдвинем гипотезу, что функция $f(x, a_0, a_1, \ldots, a_k)$ имеет вид $f(x, a_0, a_1) = a_0 + a_1 x$. В MathCAD`е поиск коэффициентов делается с помощью функций $intercept(x, y)$ и $slope(x, y)$, которые рассчитывают оценки истинных параметров.

Пусть линейная модель построена $f(x, a_0, a_1) = a_0 + a_1 x.$ В этой модели наклон a_1 представляет собой количество единиц измерения переменной y, приходящихся на одну единицу измерения переменной x. Эта величина характеризует среднюю величину изменения переменной y (положительного или отрицательного) на заданном отрезке оси x. Сдвиг a_0 представляет собой среднее значение переменной y, когда переменная x равна 0.

Возьмём некоторую точку x_0 и вычислим $y_0 = a_0 + a_1 x_0$. Заметим, что величина y_0 является случайной и меняется от выборки к выборке, следовательно её мат. ожидание в точке x_0 равно истинному значению функции. Обладая данным знанием, можно построить доверительный интервал для величины y_0.

В случае отсутствия значения дисперсии, для расчёта доверительного интервала a_0 можно использовать следующую величину S^2:

$$a0 \pm t\sqrt{S^2}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Та же ситуация применима и для доверительного интервала a_1, но с некоторыми изменениями:

$$a_1 \pm \frac{tS^2}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

И, наконец, доверительный интервал для дисперсии:

$$\left(\frac{(n-2)S^2}{\chi_{r,a}}, \frac{(n-2)S^2}{\chi_{l,a}}\right),$$

Где $\chi_{r,a}$ и $\chi_{l,a}$ рассчитываются с помощью функции «хи» - распределения $qchisq(p, d)$.

Границы доверительных интервалов в каждой точке x_0 образуют доверительную полосу, или *доверительный коридор*. Важно понимать, что эта полоса, не является доверительной областью для всей линии регрессии. Она определяет только концы доверительных интервалов для y при каждом значении x.

Доверительная область определяется как область случайных значений, которая содержит истинные значения параметров с определённой вероятностью и строится для всех значений. Данная область определяется с помощью следующих уравнений соответственно нижней и верхней границ полосы:

$$y = a_0 + a_1 x \pm 2f_\alpha \sqrt{S^2} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Примечание: функция $intercept(x, y)$ характеризует длину отрезка, отсекаемого на координатной оси прямой Y, отвечая за свободный коэффициент a_0; $slope(x, y)$ производит наклон оси и является угловым коэффициентом a_1 линейной регрессии; $qF(p, d1, d2)$ – квантиль обратного распределения Фишера, в котором $d1, d2$ – степени свободы.

Вывод: в ходе выполнения лабораторной работы, для заданной выборки:

— найдено уравнение линейной регрессии $y = a_0 + a_1 x$;

— рассчитаны оценки коэффициентов уравнения $a_0 = 3.918$ и $a_1 = -2.03$ и их доверительные интервалы $a_0(3.776, 4.060)$ и $a_1 = (-2.328, -1{,}732)$;

— построен доверительный интервал для дисперсии $Dx(0.046, 0.175)$;

— построены доверительный коридор и доверительная область регрессии.