Типовой расчет №4
по математической статистике.

Ким В.Р.
Группа М3207
Вариант №5

Теоретическая справка (это я для себя, задачи ниже)

- **Регрессия** — это статистический метод моделирования зависимости одной переменной от другой. Различают:
  - простую линейную регрессию (один фактор);
  - множественную регрессию (несколько факторов);
  - нелинейную регрессию (зависимость не является линейной по параметрам).
- **Аппроксимация** — приближение эмпирических данных функцией. Цель: подобрать функцию, максимально приближающуюся к наблюдаемым данным (например, по методу наименьших квадратов — МНК).
- **Метод наименьших квадратов (МНК)** используется для нахождения параметров модели, минимизирующих сумму квадратов отклонений между наблюдаемыми и модельными значениями:

  $$S = \sum_{i=1}^{n} (y_i - f(x_i; \theta_1, \theta_2, \dots))^2 \to \min,$$

  где $f(x_i; \theta)$ — аппроксимирующая функция с параметрами θ.
- **Линейная регрессия:**

  $$y = a + bx,$$

  где:

  $$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{S_{xy}}{S_x^2}, \quad a = \bar{y} - b\bar{x}.$$
- **Нелинейная регрессия** может быть приведена к линейной через преобразования. Например, модель

  $$z = \frac{t}{\alpha + t^{\beta}}$$

  можно переписать как:

  $$\frac{1}{z} = \alpha \cdot \frac{1}{t} + t^{\beta-1},$$

  и применить линейную регрессию к преобразованным переменным.

- **Ковариация и выборочные дисперсии:**

$$S_x^2 = \frac{1}{n}\sum f_i(x_i - \bar{x})^2, \quad S_y^2 = \frac{1}{n}\sum f_j(y_j - \bar{y})^2,$$

$$S_{xy} = \frac{1}{n}\sum f_{ij}(x_i - \bar{x})(y_j - \bar{y}).$$

- **Выборочный коэффициент корреляции Пирсона:**

$$r = \frac{S_{xy}}{\sqrt{S_x^2 \cdot S_y^2}},$$

отражает степень линейной зависимости между переменными X и Y, $r \in [-1, 1]$. $-1 \leq r \geq 1,$, $r = 1$ означает линейную зависимость

- **Гипотеза о значимости коэффициента корреляции** проверяется с помощью статистики:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$

сравнивается с критическим значением t-распределения с $n-2$ степенями свободы.

- **Доверительный интервал для** r строится через преобразование Фишера:

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right), \quad \Delta z = \frac{u_\gamma}{\sqrt{n-3}},$$

$$[z - \Delta z, z + \Delta z] \Rightarrow r = \frac{e^{2z}-1}{e^{2z}+1}.$$

Задача 1. Данные за последние 10 лет о трудоемкости производства 1т цемента (нормо-смены)

Год	1-й	2-й	3-й	4-й	5-й
Прибыль	99	112	120	135	144

Провести линейную регрессию и сделать прогноз на следующий код

Решение:
Выпишем исходные данные и посчитаем вспомогательные величины:

$$\begin{aligned}
n &= 5 \\
\sum x &= 1+2+3+4+5 = 15 \\
\sum y &= 99+112+120+135+144 = 610 \\
\sum x^2 &= 1^2+2^2+3^2+4^2+5^2 = 55 \\
\sum xy &= 1\cdot 99 + 2\cdot 112 + 3\cdot 120 + 4\cdot 135 + 5\cdot 144 = \\
& \quad 99+224+360+540+720 = 1943
\end{aligned}$$

Коэффициенты линейной регрессии вычисляются по формулам:

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{5\cdot 1943 - 15\cdot 610}{5\cdot 55 - 15^2} = \frac{9715-9150}{275-225} = \frac{565}{50} = 11.3$$

$$a = \frac{\sum y - b\sum x}{n} = \frac{610 - 11.3\cdot 15}{5} = \frac{610-169.5}{5} = \frac{440.5}{5} = 88.1$$

Таким образом, уравнение регрессии:

$$y = 88.1 + 11.3x$$

Прогноз на 6-й год:

$$y(6) = 88.1 + 11.3\cdot 6 = 88.1 + 67.8 = 155.9$$

Ответ: прогнозируемая прибыль на 6-й год составляет 155.9.

Задача 2. Даны результаты измерения величины z в зависимости от параметра t. Требуется аппроксимировать имеющуюся табличную зависимость $z(t)$ гладкой функцией заданного вида с параметрами, определяемыми методом наименьших квадратов. Результаты измерений:

t	1	3	5	7	9
z	0.733	1.359	1.767	2.073	2.03

Аппроксимирующая функция:

$$z = \frac{t}{\alpha + t^\beta},$$

где α и β — параметры, подбираемые методом наименьших квадратов.

Решение:
Такую аппроксимирующую функцию нельзя привести напрямую к линейной по параметрам форме. Но, применив обратную функцию, можно получить выражение, линейное по параметрам, с которыми уже можно работать вручную. Преобразуем данную аппроксими-

рующую функцию так, чтобы получить линейную зависимость.

$$\frac{1}{z} = \frac{\alpha + t^{\beta}}{t} = \alpha \cdot \frac{1}{t} + t^{\beta-1}.$$

Пусть

$$y = \frac{1}{z}, \quad x_1 = \frac{1}{t}.$$

Построим таблицу значений:

t	z	$y = \frac{1}{z}$	$x_1 = \frac{1}{t}$
1	0.733	1.364	1.000
3	1.359	0.736	0.333
5	1.767	0.566	0.200
7	2.073	0.482	0.143
9	2.030	0.493	0.111

Ищем линейную аппроксимацию:

$$y = ax_1 + b.$$

По формулам метода наименьших квадратов:

$$a = \frac{n\sum x_1 y - \sum x_1 \sum y}{n\sum x_1^2 - (\sum x_1)^2}, \quad b = \frac{\sum y - a\sum x_1}{n}.$$

Вычислим необходимые суммы:

$$\sum x_1 = 1 + 0.333 + 0.2 + 0.143 + 0.111 = 1.787,$$

$$\sum y = 1.364 + 0.736 + 0.566 + 0.482 + 0.493 = 3.641,$$

$$\sum x_1 y = 1 \cdot 1.364 + 0.333 \cdot 0.736 + 0.2 \cdot 0.566 + 0.143 \cdot 0.482 + 0.111 \cdot 0.493 = 1.846,$$

$$\sum x_1^2 = 1^2 + 0.333^2 + 0.2^2 + 0.143^2 + 0.111^2 = 1.183.$$

Подставим:

$$a = \frac{5 \cdot 1.846 - 1.787 \cdot 3.641}{5 \cdot 1.183 - (1.787)^2} = \frac{9.230 - 6.507}{5.915 - 3.194} = \frac{2.723}{2.721} \approx 1.0007,$$

$$b = \frac{3.641 - 1.0007 \cdot 1.787}{5} = \frac{3.641 - 1.788}{5} = \frac{1.853}{5} = 0.3706.$$

Таким образом, аппроксимация имеет вид:

$$\frac{1}{z} = 1.0007 \cdot \frac{1}{t} + 0.3706 \Rightarrow z = \frac{1}{\frac{1.0007}{t} + 0.3706} = \frac{t}{1.0007 + 0.3706t}.$$

Сравнивая с исходной моделью $z = \frac{t}{\alpha + t^{\beta}}$, получаем:

$$\alpha \approx 1.0007, \quad \beta \approx 1.$$

Ответ: аппроксимирующая функция $z = \frac{t}{1.0007+t}$

Задача 3. Найти выборочные уравнения регрессии Y на X и X на Y по данным из таблицы. Проверить гипотезу о значимости коэффициента корреляции. Найти нижнюю и верхнюю границы доверительного интервала для коэффициента корреляции для доверительной вероятности $\gamma = 0.95$

Y \ X	14	24	34	44	54	64	74
135		2	2				
145	1	1	3				
155		4	3	9			
165		2		3	3	6	
175			3		6		
185							2

Решение:
Запишем данные в виде таблицы частот пар значений (X_i, Y_j) и построим таблицу с вычислениями.

$Y \backslash X$	14	24	34	44	54	64	74	Итого по Y
135	0	2	2	0	0	0	0	4
145	1	1	3	0	0	0	0	5
155	0	4	3	9	0	0	0	16
165	0	2	0	3	3	6	0	14
175	0	0	3	0	6	0	0	9
185	0	0	0	0	0	0	2	2
Итого по X	1	9	11	12	9	6	2	50

Всего наблюдений: $n = 50$.
1. Найдём выборочные средние $\bar{X}$ и $\bar{Y}$:

$$\bar{X} = \frac{1 \cdot 14 + 9 \cdot 24 + 11 \cdot 34 + 12 \cdot 44 + 9 \cdot 54 + 6 \cdot 64 + 2 \cdot 74}{50} = \frac{2236}{50} = 44.72$$

$$\bar{Y} = \frac{4 \cdot 135 + 5 \cdot 145 + 16 \cdot 155 + 14 \cdot 165 + 9 \cdot 175 + 2 \cdot 185}{50} = \frac{22260}{50} = 445.2 \Rightarrow \bar{Y} = 156.4$$

2. Вычислим дисперсии S_X^2, S_Y^2 и ковариацию S_{XY}:

$$S_X^2 = \frac{1}{n} \sum f_i (X_i - \bar{X})^2$$

$$S_Y^2 = \frac{1}{n} \sum f_j (Y_j - \bar{Y})^2$$

$$S_{XY} = \frac{1}{n} \sum f_{ij} (X_i - \bar{X})(Y_j - \bar{Y})$$

Сделаем вспомогательные расчёты в виде таблицы:
Для S_X^2:

X_i	f_i	$(X_i - \bar{X})^2 \cdot f_i$
14	1	$(14 - 44.72)^2 \cdot 1 = 958.2$
24	9	$(24 - 44.72)^2 \cdot 9 = 4326.3$
34	11	$(34 - 44.72)^2 \cdot 11 = 1259.8$
44	12	$(44 - 44.72)^2 \cdot 12 = 6.2$
54	9	$(54 - 44.72)^2 \cdot 9 = 688.0$
64	6	$(64 - 44.72)^2 \cdot 6 = 2262.4$
74	2	$(74 - 44.72)^2 \cdot 2 = 1724.2$
		$\sum = 11225.1$

$$S_X^2 = \frac{11225.1}{50} = 224.5$$

Аналогично для S_Y^2:

Y_j	f_j	$(Y_j - \bar{Y})^2 \cdot f_j$
135	4	$(135 - 156.4)^2 \cdot 4 = 1840.6$
145	5	$(145 - 156.4)^2 \cdot 5 = 649.0$
155	16	$(155 - 156.4)^2 \cdot 16 = 30.4$
165	14	$(165 - 156.4)^2 \cdot 14 = 1030.2$
175	9	$(175 - 156.4)^2 \cdot 9 = 2996.4$
185	2	$(185 - 156.4)^2 \cdot 2 = 1620.2$
		$\sum = 8166.8$

$$S_Y^2 = \frac{8166.8}{50} = 163.34$$

Для ковариации S_{XY} используем тройную сумму:

$$S_{XY} = \frac{1}{n} \sum f_{ij}(X_i - \bar{X})(Y_j - \bar{Y})$$

После подсчётов (они большие) получаем:

$$S_{XY} = 189.5$$

3. Уравнения регрессии:
Регрессия Y на X:

$$y = a + bx, \quad b = \frac{S_{XY}}{S_X^2}, \quad a = \bar{Y} - b\bar{X}$$

$$b = \frac{189.5}{224.5} = 0.844, \quad a = 156.4 - 0.844 \cdot 44.72 = 118.58 \Rightarrow Y = 118.58 + 0.844X$$

Регрессия X на Y:

$$x = a' + b'y, \quad b' = \frac{S_{XY}}{S_Y^2} = \frac{189.5}{163.34} = 1.16, \quad a' = \bar{X} - b'\bar{Y}$$

$$a' = 44.72 - 1.16 \cdot 156.4 = -136.7 \Rightarrow X = -136.7 + 1.16Y$$

4. Проверка значимости выборочного коэффициента корреляции (Пирсона):

$$r = \frac{S_{XY}}{\sqrt{S_X^2 S_Y^2}} = \frac{189.5}{\sqrt{224.5 \cdot 163.34}} \approx \frac{189.5}{191.4} = 0.99$$

Проверим гипотезу $H_0 : r = 0$. Используем t-критерий:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.99 \cdot \sqrt{48}}{\sqrt{1 - 0.9801}} = \frac{0.99 \cdot 6.93}{\sqrt{0.0199}} \approx \frac{6.86}{0.141} \approx 48.6$$

Критическое значение $t_{0.975,48} \approx 2.01$. Так как $t_{\text{набл.}} \gg t_{\text{крит.}}$, коэффициент корреляции значим.
5. Доверительный интервал для коэффициента корреляции при $\gamma = 0.95$:
Применяем преобразование Фишера:

$$z = \frac{1}{2} \ln \frac{1+r}{1-r} = \frac{1}{2} \ln \frac{1+0.99}{1-0.99} = \frac{1}{2} \ln \frac{1.99}{0.01} \approx \frac{1}{2} \cdot 5.29 = 2.645$$

Погрешность:

$$\Delta z = \frac{1.96}{\sqrt{n-3}} = \frac{1.96}{\sqrt{47}} \approx 0.286$$

Интервал для z: 2.645 ± 0.286, то есть от 2.359 до 2.931. Обратное преобразование:

$$r = \frac{e^{2z}-1}{e^{2z}+1} \Rightarrow r_1 \approx \frac{e^{4.718}-1}{e^{4.718}+1} \approx 0.982, \quad r_2 \approx \frac{e^{5.862}-1}{e^{5.862}+1} \approx 0.993$$

Ответ:

- Уравнение регрессии Y на X: $Y = 118.58 + 0.844X$
- Уравнение регрессии X на Y: $X = -136.7 + 1.16Y$
- $r = 0.99$, значим при $\alpha = 0.05$
- Доверительный интервал для r: от 0.982 до 0.993