2 Регрессионный анализ

2.1 Введение

Часто результат испытания характеризуется не одной случайной величиной, а несколькими – *системой случайных величин* X_1, X_2,..., X_n, которую называют n-мерной случайной величиной. Каждому элементарному событию ω ставится в соответствие n действительных чисел x_1, x_2, ..., x_n, которые приняли случайные величины X_1, X_2, ..., X_n в результате испытания. Случайные величины X_1, X_2, ..., X_n могут быть как *дискретными*, так и *непрерывными*. В этой работе будем рассматривать двумерные дискретные случайные величины (X, Y). Геометрически реализацию двумерную случайную величину можно изобразить случайной точкой $A(x, y)$ (или случайным вектором (x, y) на плоскости) при этом X и Y назовем компонентами двумерного вектора (X, Y).

Полным описанием двумерной случайной величины (X, Y) является закон ее распределения. Если множество возможных значений случайной величины (X, Y) конечно, такой закон может быть задан в форме таблицы, содержащей все возможные сочетания значений каждой из одномерных случайных компонент X и Y и соответствующие им вероятности.

Пусть дискретные случайные величины X и Y получают свои значения в результате одного и того же случайного эксперимента, и принимают значения x_1, x_2, ..., x_m и y_1, y_2, ..., y_n, соответственно, где m и n некоторые целые положительные числа. Для $1 \leq i \leq m$, $1 \leq j \leq n$ определим вероятности P_{ij}, положив

$$P_{ij} = P(X = x_i, Y = y_j), \tag{27}$$

где P_{ij} равно вероятности того, что события $Y = x_i$ и $Y = y_j$ произойдут одновременно. Равенство (27) задает совместный закон распределения пары случайных величин X и Y, или закон распределения двумерного случайного вектора (X, Y). Этот закон запишем в виде следующей Таблицы 3, в которой

$$P(x_i) = \sum_{j=1}^{n} P_{ij}, P(y_j) = \sum_{i=1}^{m} P_{ij}, \tag{28}$$

т.е. $P(x_i)$ равна сумме вероятностей, расположенных в i-й строке, а $P(y_j)$ равна сумме вероятностей из j-го столбца.

Таблица 3.: Совместный закон распределения (X, Y)

XY	y_1	y_2	...	y_j	...	y_n	P_x
x_1	P_{11}	P_{12}	...	P_{1j}	...	P_{1m}	$P(x_1)$
x_2	P_{21}	P_{22}	...	P_{2j}	...	P_{2m}	$P(x_2)$
...	...	...	...	...	...	...	...
x_i	P_{i1}	P_{i2}	...	P_{ij}	...	P_{im}	$P(x_i)$
...	...	...	...	...	...	...	...
x_m	P_{m1}	P_{m2}	...	P_{mj}	...	P_{nm}	$P(x_m)$
P_y	$P(y_1)$	$P(y_2)$	...	$P(y_j)$	...	$P(y_n)$	1

С учетом несовместности событий $\{X = x_i, Y = y_j\}$ сумма вероятностей

$$\sum_{i=1}^{m}\sum_{j=1}^{n} P_{ij} = \sum_{i=1}^{m} P(x_i) = \sum_{j=1}^{n} P(y_j) = 1. \tag{29}$$

Также для данного распределения можно ввести условные вероятности.

$$\begin{aligned} P(Y = y_j|X = x_i) &= \frac{P(Y = y_j, X = x_i)}{P(x_i)} = \frac{P_{ij}}{P(x_i)}, \\ P(X = x_i|Y = y_j) &= \frac{P(Y = y_j, X = x_i)}{P(y_j)} = \frac{P_{ij}}{P(y_j)}. \end{aligned} \tag{30}$$

При изучении свойств двумерной случайной величины (X, Y) можно говорить о математических ожиданиях отдельных компонент $E(X)$ и $E(Y)$, и об их дисперсиях $D(X)$ и $D(Y)$. Однако знание характеристик изолированных компонент не позволяет делать выводы о существовании статистической связи между этими компонентами и о характере этой связи. При изучении многомерных величин дополнительно привлекают такие характеристики, как ковариация и коэффициент корреляции. Ковариация между X и Y определяется как

$$\begin{aligned} &cov(X, Y) = E\left((X - E(X)(Y - E(Y))\right) = E(XY) - E(X)E(Y) = \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n}(x_i - E(X))(y_j - E(Y))P_{ij}. \end{aligned} \tag{31}$$

Ковариационный момент характеризует линейную связь между рассматриваемыми величинами. Если случайные величины X и Y статистически независимы, т.е. $P_{ij} = P(x_i)P(y_j)$, то ковариация $cov(X, Y)$.

Коэффициент корреляции случайных величин X и Y определяется равенством

$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma_x \sigma_y}. \tag{32}$$

Коэффициент корреляции является безразмерной характеристикой, которая используется в качестве меры линейной зависимости случайных величин, причем. Очевидно, что $|\rho| \leq 1$, при этом, чем его модуль ближе к единице, тем, вообще говоря, теснее линейная зависимость между величинами. Если $|\rho| = 1$, то $Y = \beta_0 + \beta_1 X$, где β_0 и β_1 – некоторые постоянные.

2.2 Функция регрессии

Функцией регрессии (или просто *регрессией*) Y на X называется условное математическое ожидание случайной величины Y при условии, что случайная величина X приняла значение x, т.е.

$$\hat{Y} = \varphi_Y(x) = E(Y|X = x_i) = \sum_{j=1}^{n} y_j p(y_j|x_i) = \sum_{j=1}^{n} y_j \frac{P_{ij}}{P(x_i)}. \tag{33}$$

Можно показать, что регрессия Y на X обладает оптимальным свойством. Она, как функция X , наилучшим образом приближает случайную величину Y в смысле среднего квадратичного, т.е. дисперсия $E\left((Y - \hat{Y})^2\right)$ принимает минимальное значение, когда $\hat{Y} = \varphi_Y(x)$.

2.3 Линейная среднеквадратичная регрессия

Пусть двумерная случайная величина (X, Y) и её закон распределения заданы. Предположим, что X и Y связаны линейно как

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{34}$$

где ϵ_i случайная величина, причём $E(\epsilon_i|x_i) = 0$. Найдём такую линейную функцию

$$\hat{Y} = \beta_0 + \beta_1 X, \tag{35}$$

которая минимизиует отклонение $\hat{Y}$ от Y по среднему квадрату ошибки:

$$\begin{aligned} &Q(\beta_0, \beta_1) = E\left((Y - \hat{Y})^2\right) = E\left((Y - \beta_0 - \beta_1 X)^2\right) = \\ &= E(Y^2) + \beta_0^2 + \beta_1^2 E(X^2) - 2\beta_0 E(Y) - 2\beta_1 E(XY) + 2\beta_0\beta_1 E(X). \end{aligned} \tag{36}$$

Для поиска минимума $Q(\beta_0, \beta_1)$ найдём её частные производные и приравняем к нулю:

$$\begin{cases} \dfrac{\partial Q}{\beta_0} = 2\beta_0 - 2E(Y) + 2\beta_1 E(X) = 0. \\ \dfrac{\partial Q}{\beta_1} = 2\beta_1 E(X^2) - 2E(XY) + 2\beta_0 E(X) = 0. \end{cases} \tag{37}$$

Откуда получаем,

$$\begin{cases} \beta_1 = \dfrac{E(XY) - E(X)E(Y)}{E(X^2) - (E(X))^2} = \dfrac{cov(X,Y)}{D(X)} = \rho(X,Y)\dfrac{\sigma_y}{\sigma_x}, \\ \beta_0 = E(Y) - \beta_1 E(X) = E(Y) - \rho(X,Y)\dfrac{\sigma_y}{\sigma_x}E(X). \end{cases} \tag{38}$$

Таким образом, уравнение линейной регрессии может быть записано как

$$\hat{Y} = E(Y) + \rho(X,Y)\frac{\sigma_y}{\sigma_x}(X - E(X)). \tag{39}$$

Аналогично

$$\hat{X} = E(X) + \rho(X,Y)\frac{\sigma_x}{\sigma_y}(Y - E(Y)). \tag{40}$$

Пример 2.4. По закону распределения двумерного случайного вектора (X, Y) най-

	$y = 3$	$y = 5$	$P(x_i)$
$x = 2$	0.3	0.1	0.4
$x = 3$	0.4	0.4	0.6
$P(y_j)$	0.7	0.5	1

ти значения функции регрессии и функции линейной среднеквадратичной регрессии.

Найдём значения функции регрессии согласно (33). Получим $E(Y|x = 2) = 3.5$, $E(Y|x = 3) = 3.67$. Теперь найдём функцию линейной регрессии (35). Получим $E(X) = 2.6$, $E(Y) = 3.6$, $\sigma_x = 0.4899$, $\sigma_y = 0.9165$, $cov(X,Y) = 0.04$, $\beta_0 = 3.167$, $\beta_1 = 0.167$. Таким образом, $\hat{y}_i(x) = 0.167x + 3.167$, причем $\hat{y}_i(2) = 3.5$, $\hat{y}_i(2) = 3.67$, т.е. функция регрессии и функция линейной регресии совпадают, т.е. в данном случае для предсказания Y по X достаточно линейной зависимости.

2.4 Построение выборочной линии регрессии

На практике, как правило, исследователь имеет лишь случайную выборку пар значений (x_1, y_1), (x_2, y_2), ... (x_n, y_n) ограниченного объема n в предположении, что каждая пара имеет один и тот же закон

распределения. Другими словами, случайной выборкой объема n можно считать величины (x_i, y_i), $1 \leq i \leq n$, полученные в результате n независимых и одинаковых случайных экспериментов. В этом случае речь может идти об оценке функции регрессии по выборке. Наилучшей оценкой (в смысле метода наименьших квадратов) является выборочное уравнение регрессии Y на X:

$$\hat{y} = \phi(x, \beta_0, \beta_1, \beta_2, ..., \beta_m). \tag{41}$$

В случае анализа выборок пар значений, наряду с выборочными средними $\bar{X}$, $\bar{Y}$ и исправленными выборочными дисперсиями $\bar{S}_x^2$, $\bar{S}_y^2$ случайных величин X и Y, соответственно, используется несмещенная оценка ковариации $cov(X, Y)$, вычисляемая как

$$l_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{X})(y_i - \bar{Y}), \tag{42}$$

и соответствующий ей выборочный коэффициент корреляции

$$\bar{r} = \frac{l_{XY}}{\sqrt{\bar{S}_x^2 \bar{S}_y^2}}. \tag{43}$$

Существенность выборочного коэффициента корреляции, т.е. когда $|\bar{r}| \approx 1$ позволяет рассчитывать на то, что зависимость между величинами X и Y близка к линейной. Последнее возможно также визуально оценить по виду корреляционого поля точек $\{(x_i, y_i)\}$, т.е. дает основание рассчитывать на то, что теоретическое уравнение регрессии имеет вид

$$E(Y|X = x) = \beta_0 + \beta_1 x + \epsilon_i. \tag{44}$$

С учетом того, что коэффициенты β_0, β_1 неизвестны (так как неизвестна функция распределения), мы используем их соответствующие оценки $\hat{\beta}_0$ и $\hat{\beta}_1$, что даёт нам следующую выборочную линейную регрессию

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X, \tag{45}$$

где путём минимизации среднего квадрата ошибки получим, что

$$\begin{cases} \hat{\beta}_1 = \dfrac{\sum\limits_{i=1}^{n} (x_i - \bar{X})(y_i - \bar{Y})}{\sum\limits_{i=1}^{n} (x_i - \bar{X})^2} = l_{XY} \dfrac{S_{yy}}{S_{xx}}, \\ \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = \bar{Y} - l_{XY} \frac{S_{yy}}{S_{xx}} \bar{X}, \end{cases} \tag{46}$$

где $S_{xx}^2 = \sum_{i=1}^{n}(x_i - \bar{X})^2$, $S_{yy} = \sum_{i=1}^{n}(y_i - \bar{Y})^2$.

Таким образом, выборочное уравнение прямой средней квадратичной регрессии Y на X имеет вид

$$\hat{y}_i = \bar{Y} + \bar{r}\frac{S_{yy}}{S_{xx}}(x_i - \bar{X}). \tag{47}$$

При большом числе наблюдений одно и то же значение x может встретиться n_x раз, одно и то же значение y – n_y раз, а одна и та же пара чисел (x, y) может наблюдаться n_{xy} раз. Поэтому данные наблюдений группируют, т. е. подсчитывают частоты n_x, n_y и n_{xy} и все сгруппированные данные сводят в корреляционную таблицу (см. Таблицу 4).

Таблица 4.: Корреляционная таблица (X, Y)

XY	y_1	y_2	...	y_j	...	y_n	n_x
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1m}	$n(x_1)$
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2m}	$n(x_2)$
...	...	...	...	...	...	...	...
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{im}	$n(x_i)$
...	...	...	...	...	...	...	...
x_m	n_{m1}	n_{m2}	...	n_{mj}	...	n_{nm}	$n(x_m)$
n_y	$n(y_1)$	$n(y_2)$	...	$n(y_j)$	...	$n(y_n)$	n

В общем случае, x_1, x_2, ..., x_m в Таблице 4 – середины интервалов группировки по X, а y_1, y_2, ..., y_k – середины интервалов группировки по Y, n_{ij} – число точек выборки, попавших в прямоугольник с центром (x_i, y_j). Как правило, группировка осуществляется с равным шагом h_x по x и равным шагом h_y по y.

По Таблице 4 также можно определить выборочный аналог функции регрессии

$$E(Y|X = x) \approx \bar{Y}(x_i) = \sum_{j=1}^{k} y_j \frac{n_{ij}}{n(x_i)}. \tag{48}$$

Пример 2.5. Определить выборочные аналоги функции регрессии и уравнения прямой средней квадратичной регрессии Y на X для следующей корреляционной таблицы. Построим выборочное уравнение прямой средней квадратичной регрессии Y на X. Получим $\bar{X} = 24.24$,$\bar{Y} = 79.47$, $\bar{r} = 0.7901$, $S_{xx}^2 = 7.62$, $S_{yy}^2 = 15.10$. Из чего следует, что выборочная прямая регрессии задаётся как $\hat{y}_i = 1.11x_i + 52.52$.

(y_i, x_i)	18	22	26	30	n_y
70	5				5
75	7	46	1		54
80		29	72		101
85			29	8	37
90				3	3
n_x	12	75	102	11	200

x_i	18	22	26	30
$\hat{y}_i$	72.54	76.98	81.43	85.88
$\bar{Y}(x_i)$	72.92	76.93	81.37	86.36

Выбочную функцию регрессии вычислим в соответствии с (48) для значений $X = \{18, 22, 26, 30\}$. Получим следущие значения:

Как можно заметить, значения, вычисленные по уравнению выборочной регрессии и по линейной зависимости хорошо согласуются.

2.5 Приведение к линейной регрессии с помощью замены переменных

Метод наименьших квадратов (МНК), используемый для получения коэффициентов $\hat{\beta}_0$ и $\hat{\beta}_1$ линейной модели:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{49}$$

можно использовать и в более сложных случаях, когда зависимость между Y и X не является линейной. Например, зависимость

$$z = a \cdot e^{bt} \tag{50}$$

можно прологарифмировать, получив

$$\ln z = \ln a + bt, \tag{51}$$

и свести к линейной, если выполнить следующую замену:

$$\begin{cases} y = \ln z, \\ x = t, \\ \beta_0 = \ln a, \\ \beta_1 = b. \end{cases} \tag{52}$$

После вычисления β_0 и β_1 можно вернуться обратно, к исходной зависимости, получив параметры a и b.

Аналогичные действия можно выполнить, если предполагаемая зависимость имеет вид $z = a \cdot t^b$. Если зависимость $z = \frac{1}{a \cdot t + b}$, то сначала необходимо выполнить замену $y = \frac{1}{z}$. Если же зависимость $z = \frac{t}{a \cdot t + b}$, то сначала необходимо выполнить замену $y = \frac{1}{z}$, а затем замену $x = \frac{1}{t}$.

Пример 2.6. Пусть данные с некоторых измерений представлены в виде следующей таблицы. Предположим, что зависимость имеет характер $z = a \cdot e^{bt}$. Тогда сначала

t	-1	0	2	3	5
z	1.8	2	2.4	2.7	3.3
$\ln z$	0.588	0.693	0.876	0.993	1.194

вычислим значения $\ln z$. Применяя МНК, получим $\hat{\beta}_0 = 0.6876$, $\hat{\beta}_1 = 0.1006$. Теперь, с учетом (52) получим $a = 1.989$, $b = 0.101$, т.е. искомая зависимость $z = 1.989 \cdot e^{0.101t}$.

2.6 Ход выполнения работы

На языке программирования Python для заданных данных:

1. По заданному закону распределения (X, Y) сравнить функцию регрессии и функцию среднеквадратичной линейной регрессии.
2. Для заданной корреляционной таблицы определить выборочные аналоги функции регрессии и уравнения прямой средней квадратичной регрессии Y на X.
3. Для заданных данных определить, какая из зависимостей лучше всего описывает данные. Варианты зависимостей: $y = ax + b$, $z = a \cdot e^{bt}$, $z = a \cdot t^b$, $z = \frac{1}{a \cdot t + b}$ и $z = \frac{t}{a \cdot t + b}$.
4. Для выполнения работы разрешается использовать базовые вычислительные операции библиотек *numpy* или *math*, такие как сложение, вычитание, деление и т.д.