№2

2. ЛАБОРАТОРНАЯ РАБОТА №2. ЛИНЕЙНАЯ ПАРНАЯ РЕГРЕССИЯ

Цель работы – освоение методики проведения регрессионного анализа данных.

2.1 Теоретические сведения

Регрессией в математической статистике является зависимость среднего значения какой-либо величины y от некоторой другой величины x или от нескольких величин x_i.

Парной регрессией называется модель, выражающая зависимость среднего значения зависимой переменной y от одной независимой переменной x:

$$\tilde{y} = f(x), \qquad (2.1)$$

где y – зависимая переменная (результативный признак); x – независимая переменная (признак–фактор).

Парная регрессия применяется в ситуациях, когда имеется доминирующий фактор, обуславливающий большую долю изменения изучаемой объясняемой переменной.

Регрессионный анализ – раздел математической статистики, объединяющий практические методы исследования регрессионной зависимости между величинами по статистическим данным.

Если корреляционный анализ позволяет выявить сам факт наличия зависимости между параметрами, изменяющимися случайным образом, то регрессионный анализ позволяет установить (при определенных предположениях) вид этой зависимости. В частности, по имеющимся наблюдениям значений изучаемых параметров, исходя из априори известной по каким-либо соображениям структуры закона взаимосвязи между этими параметрами, регрессионный анализ позволяет определить числовые значения коэффициентов, входящих в описание этого закона.

Цель регрессионного анализа состоит в определении общего вида уравнения регрессии, построении оценок неизвестных параметров, входящих в уравнение регрессии, и проверке статистических гипотез о регрессии. При изучении связи между двумя величинами по результатам наблюдений $(x_i, y_i), \ldots, (x_n, y_n)$ в соответствии с теорией регрессии предполагается, что одна из них y имеет некоторое распределение вероятностей при фиксированном значении x другой.

В зависимости от формы связи между переменными различают ***линейную*** и ***нелинейную*** регрессию.

Наиболее простым является случай, когда регрессия y по x линейна.

Уравнение линейной парной регрессии представляют в виде:

$$\tilde{y} = a + bx \tag{2.2}$$

где y – результирующий признак;

x – факторный признак;

a и b – числовые параметры (коэффициенты) уравнения.

Коэффициенты a и b в уравнении регрессии называются ***коэффициентами регрессии.***

Построение уравнения регрессии осуществляется в два этапа (предполагает последовательное решение двух задач):

- спецификация модели (определение вида аналитической зависимости);
- оценка параметров выбранной модели.

Применяются три основных метода выбора вида аналитической зависимости:

- графический (на основе анализа поля корреляций);
- аналитический (исходя из теории изучаемой взаимосвязи);
- экспериментальный (путем сравнения величины остаточной дисперсии $D_{ост}$ или средней ошибки аппроксимации A, рассчитанных для различных моделей регрессии - метод перебора).

Для нахождения уравнения линейной парной регрессии необходимо определить значения коэффициентов a и b в уравнении (2.2). Одним из наиболее распространенных методов, который позволяет вычислить значения коэффициентов, является метод наименьших квадратов.

Основная идея метода наименьших квадратов заключается в минимизации квадратов ошибок (расстояний) от экспериментальных точек до точек, расположенных на теоретической прямой линии.

Для определения параметров a, b методом наименьших квадратов необходимо решить следующую систему нормальных уравнений:

$$\begin{cases} na + b\sum x = \sum y, \\ a\sum x + b\sum x^2 = \sum yx \end{cases} \tag{2.3}$$

В результате решения системы (2.3) получим:

$$b = \frac{\operatorname{cov}(x;y)}{\sigma_x^2} = \frac{\overline{y \cdot x} - \overline{y} \cdot \overline{x}}{\overline{x^2} - \overline{x}^2}; \qquad a = \overline{y} - b \cdot \overline{x} \tag{2.4}$$

где $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ – среднее значение фактора x;

$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ – среднее значение результативной переменной y;

$\overline{y \cdot x} = \frac{1}{n}\sum_{i=1}^{n} y_i \cdot x_i$ – среднее значение произведения переменных x и y;

$\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$ – среднее значение квадрата переменной x;

$\mathrm{cov}(x; y) = \overline{x \cdot y} - \overline{x} \cdot \overline{y}$ – ковариация переменных x и y;

$\sigma_x^2 = D_x = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2 = \overline{x^2} - \overline{x}^2$ – дисперсия переменной x.

Коэффициент регрессии b показывает, на сколько единиц в среднем по совокупности изменится результирующая переменная y, если факторная переменная x увеличится на одну единицу.

Получив уравнение регрессии, необходимо провести оценку его значимости.

Проверка значимости уравнения регрессии предполагает выяснение ответов на два важных вопроса:

- соответствует ли математическая модель, выражающая зависимость между переменными, экспериментальным данным ?;
- достаточно ли включенных в уравнение объясняющих переменных для описания зависимой переменной ?.

Точность построенной модели регрессии можно оценить по средней квадратической ошибке:

$$\varepsilon_{\text{кв}} = \sqrt{\frac{\sum_{i=1}^{n} \left(\tilde{y}_i - y_i\right)^2}{n}} \qquad (2.5)$$

Для оценки качества модели используют среднюю ошибку аппроксимации, которая представляет собой среднее относительное отклонение расчетных значений от наблюдаемых:

$$\overline{A} = \frac{1}{n}\sum_{i=1}^{n} \left| \frac{y_i - \tilde{y}_i}{y_i} \right| \cdot 100\% \qquad (2.6)$$

Построенное уравнение регрессии можно считать удовлетворительным, если величина $\overline{A}$ не превышает 8–10 %.

В основе проверки значимости регрессии лежит идея разложения дисперсии (разброса) результативного признака на факторную $D_{\text{факт}}$ и остаточную $D_{\text{ост}}$ дисперсии, т.е. объясненную (за счет независимых факторов) часть дисперсии и часть, оставшуюся необъясненной в рамках данной модели:

$D_{\text{общ}} = \sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2$ - общая дисперсия;

$D_{\text{факт}} = \sum_{i=1}^{n} (\tilde{y}_i - \bar{y})^2$ - факторная (объясненная) дисперсия;

$D_{\text{ост}} = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$ - остаточная (необъясненная) дисперсия.

Долю дисперсии одного признака, объясняемую влиянием другого определяет коэффициент детерминации:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\tilde{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \qquad (2.7)$$

где n – количество наблюдений; x_i, y_i – данные наблюдений; $\bar{y}$ - среднее значение переменной y; $\tilde{y}_i$ – расчетные значения переменной y, вычисленные по уравнению регрессии.

Из определения коэффициента детерминации следует, что он принимает значения в диапазоне от 0 до 100 %.

В силу определения R^2: $0 \le R^2 \le 1$.

Чем ближе величина R^2 к единице, тем лучше уравнение регрессии $\tilde{y} = f(x)$ согласуется с данными наблюдений. При R^2 = 1 соотношение $\tilde{y}_i = f(x_i)$ выполняется для всех наблюдений, т.е. зависимость $\tilde{y} = f(x)$ является функциональной.

Величина R^2 показывает, какая доля общей дисперсии (вариации) результативного признака y объясняется уравнением регрессии. Например, значение R^2 = 0,8 означает, что уравнение регрессии объясняет 80 % общей дисперсии (вариации) результативного признака y. Таким образом, по величине R^2 можно судить о том, насколько хорошо модель подходит под исходные данные.

Сопоставляя факторную $D_{\text{факт}}$ и остаточную $D_{\text{ост}}$ дисперсии в расчете на одну степень свободы, получим величину F - отношения (F- критерия):

$$F = \frac{D_{\text{факт}}}{D_{\text{ост}}}, \qquad (2.8)$$

где F-критерий для проверки нулевой гипотезы H_0: $D_{\text{факт}} = D_{\text{ост}}$.

Если нулевая гипотеза справедлива, то по сути это означает, что на рузультативный признак y в равной степени влияют и независимамая (факторная) переменная x и необъясненные факторы – ***уравнение регрессии не значимо.***

Чтобы уравнение регрессии было значимым необходимо, чтобы факторная дисперсия превышала остаточную в несколько раз. Следовательно, необходимо произвести статистическую проверку полученного уравнения – попытаться опровергнуть гипотезу H_0.

Оценка статистической значимости уравнения регрессии в целом осуществляется с помощью F-критерия Фишера аналогично проверке статистической значимости коэффициента детерминации для парной линейной регрессии.

Расчетное значение F-критерия Фишера можно определить по выражению:

$$F_{\text{р}} = \frac{R^2}{1-R^2} \cdot \frac{n-p-1}{p}, \tag{2.9}$$

где n – количество наблюдений; p – число параметров при факторных переменных. Для парной линейной регрессии $p = 1$.

Расчетное значение $F_{\text{р}}$ необходимо сравнить с табличным $F_{\text{табл}}$ (определяется по таблице П1.2 приложения при необходимом уровне значимости α и заданном числе степеней свободы $k_1 = p$, $k_2 = n–p–1$):

$$F_{\text{р}} \geq F_{\text{табл}}, \tag{2.10}$$

Если условие (2.10) выполняется, то нулевая гипотеза H_0 о статистической незначимости уравнения регрессии отвергается и уравнение считается статистически значимым.

Статистическая значимость коэффициентов уравнения регрессии, также как и для коэффициента корреляции, определяется через t-критерий Стьюдента. Выдвигается гипотеза H_0 о случайной природе коэффициентов, т.е. о незначимом их отличии от нуля. Наблюдаемые значения t-критерия рассчитываются по формулам:

$$t_b = \frac{b}{m_b};\ t_a = \frac{a}{m_a};\ t_r = \frac{|r_{xy}|}{m_r}, \tag{2.11}$$

где m_b, m_a, m_r – случайные ошибки параметров линейной регрессии и коэффициента корреляции.

Для линейной парной регрессии выполняется равенство $t_b = t_r = \sqrt{F}$, поэтому проверки гипотез о значимости коэффициента регрессии при факторе и коэффициента корреляции равносильны проверке гипотезы о статистической значимости уравнения регрессии в целом.

В общем случае, случайные ошибки линейной регрессии рассчитываются по формулам:

$$m_r = \sqrt{\frac{1-r_{xy}^2}{n-2}};\ m_b = \frac{S_{\text{ост}}}{\sigma_x\sqrt{n}};\ m_a = S_{\text{ост}} \frac{\sqrt{\sum_{i=1}^{n} x_i^2}}{n\sigma_x}, \tag{2.12}$$

где $S_{ост}^2$ – остаточная дисперсия на одну степень свободы:

$$S_{\text{ост}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{n-2}} \tag{2.13}$$

Табличное (критическое) значение t-статистики находят по таблицам распределения t-Стьюдента при требуемом уровне значимости α и числе степеней свободы $k = n - 2$ (по таблице П1.1 приложения). Если $t_т < t_р$, то H_0 отклоняется, т.е. коэффициенты регрессии не случайно отличаются от нуля и сформировались под влиянием систематически действующего фактора.

Проверка значимости оценок параметров не дает информации о том, насколько эти оценки могут отличаться от точных значений. Ответ на этот вопрос дает построение доверительных интервалов. Под доверительным интервалом понимаются пределы, в которых лежит точное значение определяемого показателя с заданной вероятностью (P=1–α). Доверительные интервалы для точных значений параметров $\tilde{a}$ и $\tilde{b}$ уравнения линейной регрессии определяются соотношениями:

$$\begin{aligned} a - t_т \cdot m_a < \tilde{a} < a + t_т \cdot m_a; \\ b - t_т \cdot m_b < \tilde{b} < b + t_т \cdot m_b; \end{aligned} \qquad (2.14)$$

Величина $t_т$ представляет собой табличное значение t-критерия Стьюдента при уровне значимости α и числе степеней свободы n–2.

Если в границы доверительного интервала попадает ноль, т. е. нижняя граница отрицательна, а верхняя положительна, то оцениваемый параметр принимается равным нулю, так как он не может одновременно принимать и положительное, и отрицательное значения.

После получения уравнения регрессии часто возникает необходимость выполнения точечного и интервального прогнозов.

Точечный прогноз заключается в получении прогнозного значения $y_{пр}$, которое определяется путем подстановки в уравнение регрессии $\tilde{y} = a + bx$ соответствующего (прогнозного) значения $x_{пр}$:

$$y_{пр} = a + b \cdot x_{пр} \qquad (2.15)$$

Выполнение интервального прогноза заключается в получении доверительного интервала прогноза, т. е. нижней $y_{пр.min}$ и верхней $y_{пр.max}$ границ интервала, содержащего точную величину для прогнозного значения $y_{пр}$ с заданной вероятностью:

$$y_{пр.min} < y_{пр} < y_{пр.max} \qquad (2.16)$$

При построении доверительного интервала прогноза используется стандартная ошибка индивидуального значения прогноза $m_{y\_пр}$, связанная с дисперсией ошибки прогноза соотношением:

$$m_{y\_пр} = S_{ост} \cdot \sqrt{1 + \frac{1}{n} + \frac{\left(x_{пр} - \overline{x}\right)^2}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}} \qquad (2.17)$$

Тогда доверительный интервал для индивидуального значения прогноза $y_{пр}$ определяется соотношением:

$$y_{пр} - t_т \cdot m_{y\_пр} < y_{пр} < y_{пр} + t_т \cdot m_{y\_пр} \qquad (2.18)$$

2.2. Методические указания

По данным таблицы 1.2 (лабораторная работа №1), для соответствующего варианта задания необходимо:

1. С использованием метода наименьших квадратов получить уравнение линейной парной регрессии.

Для выполнения вычислений рекомендуется использовать таблицу 2.1.

Таблица 2.1 – Таблица для расчета коэффициентов уравнения регрессии

n	x	y	x^2	y^2	$x \cdot y$	$\tilde{y}$	$(x-\overline{x})^2$	$(y-\overline{y})^2$	$(\tilde{y}-\overline{y})^2$	$(y-\tilde{y})^2$
1 2 3 4 5 …. n										
Сумма Среднее			- -	- -	- -	- -	-			

2. Определить среднюю квадратическую ошибку уравнения регрессии.

3. Определить среднюю ошибку аппроксимации.

4. Найти коэффициент детерминации.

4. Проверить значимость уравнения регрессии при уровне значимости α=0,05.

5. Проверить значимость коэффициентов линейной регрессии и построить доверительные интервалы для точных значений параметров a и b уравнения линейной регрессии с уровнем значимости 0.05.

6. Построить точечный и интервальный прогноз для значения $x_{пр}=0.7 \cdot x_{max}$ по уравнению линейной регрессии с уровнем значимости 0.05.

7. Представить результаты моделирования в графическом виде.

2.3. Требования к отчету

Отчет по лабораторной работе должен содержать:

1. Цель работы
2. Результаты проведенных вычислений
3. Выводы

2.4. Контрольные вопросы

1. Каково назначение регрессионного анализа?
2. Что такое уравнение регрессии?
3. Какие виды регрессии различают?
4. В чем заключается задача построения регрессионной зависимости?
5. Для чего применяется F-критерий Фишера, как он вычисляется?
6. Как вычисляется и что показывает коэффициент детерминации?
7. Как проверяется значимость уравнения регрессии?
8. Как проверяется значимость коэффициентов уравнения регрессии?
9. Понятие доверительного интервала для коэффициентов регрессии.
10. Понятие точечного и интервального прогноза по уравнению линейной регрессии.