Урок 11. Корреляционный анализ

Хакимов Р.И. + ChatGPT

Введение в корреляцию

Корреляция – это статистический метод, используемый для измерения и анализа степени и направления взаимосвязи между двумя количественными переменными.

Основная цель корреляционного анализа – определить, насколько сильно и в каком направлении одна переменная связана с другой.

Корреляция используется в различных областях, таких как экономика, медицина, социология и психология, для изучения взаимосвязей между переменными и для построения предсказательных моделей.

Основные концепции корреляции

Корреляционный коэффициент

Это мера силы и направления линейной зависимости между двумя переменными.

Значения корреляционного коэффициента (r) варьируются от -1 до 1:

- $r = 1$: Идеальная положительная линейная зависимость.
- $r = -1$: Идеальная отрицательная линейная зависимость.
- $r = 0$: Отсутствие линейной зависимости.

Типы корреляции:

Положительная корреляция: Если одна переменная увеличивается, то другая переменная также увеличивается.

Отрицательная корреляция: Если одна переменная увеличивается, то другая переменная уменьшается.

Корреляция и линейная зависимость:

Корреляция измеряет только линейную зависимость между переменными. Если зависимость нелинейная, корреляция может не отражать истинную связь.

Коэффициент корреляции Пирсона

Коэффициент корреляции Пирсона измеряет линейную зависимость между двумя переменными. Он может принимать значения от -1 (идеальная обратная корреляция) до 1 (идеальная прямая корреляция), а 0 указывает на отсутствие линейной зависимости.
Формула для вычисления коэффициента корреляции Пирсона (r):

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

где X_i и Y_i — значения двух переменных, $\bar{X}$ и $\bar{Y}$ — средние значения переменных.
Этот коэффициент применяется, когда данные нормально распределены и связь между переменными линейная.

Коэффициент корреляции Пирсона

Шаги вычисления:

1. Найдите средние значения $\bar{x}$ и $\bar{y}$ для каждой переменной x и y:

$$\bar{x} = \frac{\sum x_i}{n}, \quad \bar{y} = \frac{\sum y_i}{n}$$

2. Вычислите отклонения каждого значения от среднего для обеих переменных: $(x_i - \bar{x})$ и $(y_i - \bar{y})$.
3. Найдите произведение отклонений для каждой пары значений $(x_i - \bar{x})(y_i - \bar{y})$.
4. Просуммируйте произведения отклонений:

$$\sum (x_i - \bar{x})(y_i - \bar{y})$$

5. Найдите квадрат отклонений для каждой переменной: $(x_i - \bar{x})^2$ и $(y_i - \bar{y})^2$.

Коэффициент корреляции Пирсона

Шаги вычисления:

6. Просуммируйте квадраты отклонений для обеих переменных:

$$\sum(x_i - \bar{x})^2 \quad \text{и} \quad \sum(y_i - \bar{y})^2$$

7. Подставьте значения в формулу для расчёта коэффициента Пирсона:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Интерпретация:

- $r = 1$ — идеальная положительная корреляция,
- $r = -1$ — идеальная отрицательная корреляция,
- $r = 0$ — отсутствует линейная зависимость.

Коэффициент корреляции Пирсона. Пример

Предположим, вы хотите исследовать зависимость между количеством часов, проведенных за изучением, и оценками на экзамене. Вы собрали данные для группы студентов:

Часы изучения	Оценка на экзамене
1	55
2	60
3	65
4	70
5	75

Для вычисления коэффициента корреляции Пирсона:

1. Вычислите средние значения для двух переменных:

$$\bar{X} = \frac{1+2+3+4+5}{5} = 3$$

$$\bar{Y} = \frac{55+60+65+70+75}{5} = 65$$

Коэффициент корреляции Пирсона. Пример

2. Вычислите сумму произведений отклонений от средних:

$$\sum(X_i - \bar{X})(Y_i - \bar{Y})$$

Для каждой пары (X, Y):

(1 - 3)(55 - 65) = 20
(2 - 3)(60 - 65) = 5
(3 - 3)(65 - 65) = 0
(4 - 3)(70 - 65) = 5
(5 - 3)(75 - 65) = 20
Сумма = 50

3. Вычислите сумму квадратов отклонений:

Сумма квадратов отклонений X: $\sum(X_i - \bar{X})^2 = 4 + 1 + 0 + 1 + 4 = 10.$
Сумма квадратов отклонений Y: $\sum(Y_i - \bar{Y})^2 = 100 + 25 + 0 + 25 + 100 = 250$

Коэффициент корреляции Пирсона. Пример

4. Расчет коэффициента корреляции:

$$r = \frac{50}{\sqrt{10 \times 250}} = \frac{50}{\sqrt{2500}} = \frac{50}{50} = 1$$

Вывод: Коэффициент корреляции $r = 1$ указывает на идеальную положительную линейную зависимость между часами изучения и оценками на экзамене.

Коэффициент корреляции Спирмена

Коэффициент корреляции Спирмена – это непараметрическая мера статистической зависимости между двумя переменными. Он основан на рангах значений, а не на самих значениях.

Используется для измерения силы и направления монотонной (не обязательно линейной) зависимости между переменными. Полезен, когда данные не соответствуют нормальному распределению.

Формула для вычисления коэффициента корреляции Спирмена:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

где:

- r_s — коэффициент корреляции Спирмена,
- d_i — разница между рангами соответствующих значений двух переменных для каждой пары наблюдений,
- n — количество наблюдений.

Коэффициент корреляции Спирмена

Шаги вычисления:

1. Упорядочьте значения каждой переменной в виде рангов.
2. Найдите разницу рангов для каждой пары значений.
3. Возведите разницу рангов d_i в квадрат.
4. Просуммируйте квадраты разностей рангов.
5. Подставьте полученные значения в формулу.

Коэффициент r_s принимает значения от -1 до 1:

$r_s = 1$ указывает на полную положительную зависимость (идеальная прямая связь),
$r_s = -1$ указывает на полную отрицательную зависимость (идеальная обратная связь),
$r_s = 0$ говорит об отсутствии корреляции.

Коэффициент корреляции Спирмена. Пример

Рассмотрим следующий набор данных, в котором представлены оценки студентов по математике и физике:

Студент	Математика (X)	Физика (Y)
1	85	80
2	90	85
3	78	70
4	92	90
5	88	82

Коэффициент корреляции Спирмена. Пример

Шаг 1: Присвоение рангов

Сначала мы присваиваем ранги значениям в обеих колонках. Если два значения одинаковы, им присваивается средний ранг:

Студент	Математика (X)	Ранг X	Физика (Y)	Ранг Y
1	85	2	83	3
2	90	4	85	4
3	72	1	75	1
4	92	5	90	5
5	88	3	81	2

Коэффициент корреляции Спирмена. Пример

Шаг 2: Вычисление разностей рангов

Теперь вычисляем разности рангов d_i:

Студент	Ранг X	Ранг Y	d_i = Ранг X − Ранг Y	d_i^2
1	2	3	-1	1
2	4	4	0	0
3	1	1	0	0
4	5	5	0	0
5	3	2	1	1

Шаг 3: Подсчет суммы квадратов разностей

Теперь подсчитываем сумму квадратов разностей:

$$\sum d_i^2 = 1 + 0 + 0 + 0 + 1 = 2$$

Коэффициент корреляции Спирмена. Пример

Шаг 4: Подстановка в формулу

Теперь можем подставить значения в формулу для вычисления коэффициента корреляции Спирмена:

$$\rho_s = 1 - \frac{6 \times 2}{5(5^2 - 1)} = 1 - \frac{12}{5 \times 24} = 1 - \frac{12}{120} = 1 - 0.1 = 0.9$$

Интерпретация результата

Коэффициент корреляции Спирмена равен 0.9, что указывает на положительную монотонную зависимость между оценками студентов по математике и физике.

Графическое представление корреляции

Графическое представление корреляции позволяет наглядно увидеть взаимосвязь между двумя количественными переменными. Это помогает в интерпретации данных и понимании природы их связи. Основные методы графического представления корреляции включают диаграмму рассеяния и линии тренда.

Графическое представление корреляции

Диаграмма рассеяния (scatter plot) — это графическое представление данных, где каждая точка на плоскости соответствует одной паре значений переменных. Она позволяет увидеть, как значения одной переменной соотносятся со значениями другой переменной и выявить наличие линейной или нелинейной зависимости.

Структура диаграммы рассеяния:

- Ось X: значения первой переменной (например, количество часов изучения).
- Ось Y: значения второй переменной (например, оценка на экзамене).

Графическое представление корреляции

Как интерпретировать диаграмму рассеяния:

1. Положительная линейная зависимость: Если точки диаграммы располагаются вдоль восходящей линии, это указывает на положительную корреляцию. Например, увеличение часов изучения связано с увеличением оценок на экзамене.

2. Отрицательная линейная зависимость: Если точки диаграммы располагаются вдоль нисходящей линии, это указывает на отрицательную корреляцию. Например, увеличение температуры связано с уменьшением времени, проведенного на улице.

3. Отсутствие явной зависимости: Если точки распределены случайным образом, это может указывать на отсутствие или слабую корреляцию между переменными.

4. Нелинейные зависимости: Если точки формируют кривую, это может указывать на нелинейную зависимость между переменными.

Графическое представление корреляции

Линия тренда (или линия регрессии) добавляется к диаграмме рассеяния, чтобы лучше визуализировать линейную зависимость между переменными. Она показывает направление и силу линейной связи между переменными.

Как построить линию тренда:

1. Выполните линейную регрессию для определения линии тренда.
2. Постройте линию на диаграмме, используя уравнение регрессии. Обычно эта линия имеет вид $Y = a + bX$, где a — свободный член, а b — коэффициент наклона (угол наклона).

Интерпретация:

- Если линия тренда имеет положительный наклон, это указывает на положительную корреляцию: больше часов тренировки связано с лучшими результатами в соревнованиях.
- Если линия тренда почти горизонтальна, это указывает на отсутствие значительной корреляции.

Интерпретация корреляционного анализа

Интерпретация корреляционного анализа включает в себя понимание и объяснение результатов, полученных при вычислении коэффициента корреляции, а также выводы, которые можно сделать на основе графического представления данных.
Далее некоторые ключевые аспекты интерпретации корреляционного анализа...

Интерпретация корреляционного анализа

1. Коэффициент корреляции

Коэффициент корреляции (r) измеряет силу и направление линейной зависимости между двумя количественными переменными. Он может принимать значения от -1 до 1:

$r = 1$: Идеальная положительная линейная зависимость. Все точки на диаграмме рассеяния лежат на прямой линии, направленной вверх.

$r = -1$: Идеальная отрицательная линейная зависимость. Все точки на диаграмме рассеяния лежат на прямой линии, направленной вниз.

$r = 0$: Отсутствие линейной зависимости. Точки на диаграмме рассеяния распределены случайным образом.

Сила корреляции:

$0 < |r| < 0.3$: Слабая корреляция.

$0.3 \leq |r| < 0.7$: Умеренная корреляция.

$0.7 \leq |r| \leq 1$: Сильная корреляция.

Интерпретация корреляционного анализа

1. Коэффициент корреляции

Направление корреляции:

$r > 0$: Положительная корреляция. Когда одна переменная увеличивается, другая также увеличивается.

$r < 0$: Отрицательная корреляция. Когда одна переменная увеличивается, другая уменьшается.

Пример интерпретации:

Если коэффициент корреляции между количеством часов работы и производительностью составляет 0.85, это указывает на сильную положительную корреляцию. Это может значить, что увеличение количества часов работы связано с увеличением производительности.

Интерпретация корреляционного анализа

2. Графическое представление

Диаграмма рассеяния:

- На диаграмме рассеяния можно визуально оценить характер связи между переменными.
- Если точки на диаграмме располагаются вдоль прямой линии, это подтверждает линейную зависимость.
- Распределение точек по диаграмме помогает понять, насколько хорошо одна переменная предсказывает другую.

Линия тренда:

- Линия тренда показывает направление и силу связи. Чем лучше линия тренда соответствует точкам данных, тем сильнее корреляция.
- Если линия тренда почти горизонтальная, это указывает на слабую корреляцию или её отсутствие.

Интерпретация корреляционного анализа

3. Корреляция vs. Причинно-следственная связь

Важно помнить, что корреляция не подразумевает причинно-следственной связи. Даже если две переменные сильно коррелируют, это не означает, что одна переменная вызывает изменения в другой. Корреляция просто указывает на наличие взаимосвязи.

Пример:

Высокий коэффициент корреляции между количеством выпитого кофе и уровнем энергии не означает, что кофе непосредственно повышает уровень энергии. Могут быть другие факторы, такие как общие привычки или уровень стресса, которые влияют на оба параметра.

Интерпретация корреляционного анализа

4. Проверка значимости

Проверка значимости корреляции помогает определить, является ли наблюдаемая корреляция статистически значимой, а не случайной. Это обычно делается с помощью p-значения:

$p < 0.05$: Корреляция статистически значима, и есть основание полагать, что связь не является случайной.

$p \geq 0.05$: Корреляция может быть случайной, и нет достаточных доказательств значимой связи.

Заключение

Интерпретация корреляционного анализа требует понимания как количественных (коэффициент корреляции, значимость), так и качественных аспектов (графическое представление). Это позволяет делать обоснованные выводы о взаимосвязях между переменными, а также формировать гипотезы для дальнейшего анализа и исследования.

Ограничения корреляционного анализа

Корреляционный анализ предоставляет полезную информацию о взаимосвязи между двумя переменными, но он имеет свои ограничения. Вот основные из них:

1. Корреляция не подразумевает причинно-следственную связь
Корреляция показывает, что две переменные связаны, но не указывает на причину этой связи. Даже при наличии сильной корреляции между переменными нельзя утверждать, что одна переменная вызывает изменение другой.
Пример: Если наблюдается высокая корреляция между количеством выпитого кофе и уровнем энергии, это не означает, что кофе непосредственно повышает уровень энергии. Возможно, есть другие факторы, влияющие на обе переменные, такие как общий уровень стресса или режим сна.

Ограничения корреляционного анализа

2. Корреляция измеряет только линейную зависимость
Коэффициент корреляции Пирсона измеряет только линейную зависимость между переменными. Если связь между переменными нелинейная, корреляция может быть низкой, даже если существует сильная зависимость.
Пример: Если переменные связаны квадратичной зависимостью (например, $Y = X^2$), коэффициент корреляции Пирсона может не отражать этого, и значение r может быть близким к нулю.

3. Чувствительность к выбросам
Корреляция сильно подвержена влиянию выбросов или аномальных значений. Один или несколько выбросов могут существенно изменить значение коэффициента корреляции.
Пример: Если в наборе данных есть несколько точек, которые значительно отклоняются от основной массы данных, это может искажать результаты корреляционного анализа и приводить к неверным выводам.

Ограничения корреляционного анализа

4. Корреляция может быть случайной

При небольших объемах данных корреляция может быть случайной. Небольшие выборки могут показывать корреляцию, которая в реальности отсутствует.

Пример: При анализе данных, собранных из небольшой группы людей, можно обнаружить корреляцию, которая не будет наблюдаться в более крупной выборке или в других исследованиях.

5. Не учитывает другие переменные

Корреляция не учитывает влияние других переменных. Если есть третья переменная, которая влияет на обе изучаемые переменные, это может исказить результаты корреляционного анализа.

Пример: Если исследуется связь между количеством часов, проведенных за изучением, и оценками на экзамене, не учитывая, что студенты могут также различаться по уровню мотивации или предыдущему опыту, результат может быть неполным.

Ограничения корреляционного анализа

6. Не показывает характер зависимости

Коэффициент корреляции не показывает, насколько сильна зависимость между переменными. Например, корреляция 0.8 и 0.9 могут указывать на сильную зависимость, но разница в значении не всегда отражает практическое значение.

Пример: Хотя корреляция 0.8 может указывать на сильную связь, интерпретация и оценка того, насколько эта связь значима, требует дополнительного анализа и контекста.

Заключение

Корреляционный анализ — это мощный инструмент для изучения взаимосвязей между переменными, но его ограничения требуют внимательного подхода при интерпретации результатов. Важно помнить, что корреляция не заменяет более глубокий анализ причинно-следственных связей и не учитывает все возможные факторы, влияющие на переменные. Использование корреляционного анализа в сочетании с другими методами, такими как регрессионный анализ и контроль дополнительных переменных, может помочь получить более полное представление о данных.