Необходимые условия практического использования корреляционно-регрессионного анализа

- **Однородность изучаемой статистической совокупности.**
- **Достаточно большой объем совокупности (условие действия закона больших чисел). Число единиц совокупности должно быть в 5 – 6 (идеально в 10) раз больше числа факторов, влияние которых предполагается оценить.**
- **Устойчивость влияния факторов, включаемых в анализ.**
- **Признаки-факторы должны иметь количественную оценку, что необходимо для построения уравнения регрессии.**
- **Отсутствие тесной линейной зависимости между факторами (коллинеарности, мультиколлинеарности).**
- **Независимость наблюдений.**
- **Желательно, чтобы распределение единиц изучаемой совокупности соответствовало закону нормального распределения.**
- **Прежде, чем воспользоваться сложными вычислительными процедурами корреляционно-регрессионного анализа, полезно на основе фактических данных убедиться в наличии корреляционной связи между интересующими исследователя признаками, определить ее характер и направленность.**

Корреляционный и регрессионный анализ данных.
Множественный регрессионный анализ.

- **Методы выявления корреляционной зависимости**

  \- Построение и анализ параллельных рядов. При этом строится ранжированный ряд значений факторного признака и параллельно – ряд соответствующих значений признака-результата. По согласованному или несогласованному изменению значений фактора и результата судят о наличии либо отсутствии зависимости.

  \- Построение и анализ групповых таблиц. Групповая таблица строится по правилам аналитической группировки. В качестве группировочного признака используется факторный признак. По каждой из выделенных групп рассчитывается среднее значение результативного признака. Наличие закономерности в изменении средних величин зависимой переменной будет свидетельствовать о присутствии корреляционной связи.

Корреляционный и регрессионный анализ данных.
Множественный регрессионный анализ.

- Построение и анализ корреляционных таблиц. В отличие от групповых, построение корреляционных таблиц предполагает группировку данных и по признаку-фактору, и по признаку-результату. На пересечении строк и столбцов проставляют частоты, т.е. число единиц совокупности с данным сочетанием уровней изучаемых признаков. Характер расположения частот на поле таблицы позволяет выдвинуть предположение о наличии и направлении зависимости между признаками.

- Графический метод. В прямоугольной системе координат по оси абсцисс откладываются значения признака-фактора, а по оси ординат – значения результативного признака. Точки на графике соответствуют единицам совокупности с конкретными сочетаниями значений признаков. Получаемый точечный график называют "полем корреляции". По расположению точек на графике судят о наличии или отсутствии зависимости, а также о направлении и степени тесноты корреляционной связи.

Корреляционный анализ

- Первой и простейшей характеристикой тесноты связи является линейный коэффициент парной корреляции.
- Показатели корреляции основаны на оценке сопряженной вариации изучаемых признаков. Парный коэффициент корреляции *(r)* – это нормированный коэффициент ковариации. Ковариация, являясь мерой взаимосвязи двух переменных, рассчитывается как средняя величина произведения отклонений индивидуальных значений анализируемых признаков от их средних значений:

$$Cov(y,x)=\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x}) \qquad (1)$$

Парный коэффициент корреляции:

$$r = \frac{\sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right)}{n} \qquad (2)$$

где n – число единиц в статистической совокупности, σy – среднее квадратическое отклонение признака-результата; σx – среднее квадратическое отклонение признака-фактора.

Линейный коэффициент корреляции Пирсона:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n\sigma_x\sigma_y} \qquad (3)$$

или

$$r = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sigma_x\sigma_y} \qquad (4)$$

Коэффициент корреляции изменяется в пределах

$$0 \le |r| \le 1 \qquad (5)$$

Если $r = 0$, линейная связь между изучаемыми признаками отсутствует. Если $|r| = 1$, связь функциональная, т.е. значение зависимой переменной полностью определяется независимой переменной. Положительное значение коэффициента свидетельствует о прямой зависимости между признаками, отрицательная – об обратной.

Парный коэффициент корреляции – это симметричная характеристика, т.е. $r_{yx} = r_{xy}$.

Значение r отражает только степень тесноты корреляционной связи между изучаемыми признаками, но не свидетельствует о причинно-следственной зависимости между ними.

Обоснование наличия причинно-следственной связи между признаками опирается на анализ природы изучаемого явления.

- Квадрат коэффициента корреляции (r^2) называется коэффициентом детерминации. Его значение изменяется в пределах от 0 до 1, и означает долю вариации результативного признака, обусловленную вариацией признака-фактора.

- Парный коэффициент корреляции достаточно точно оценивает тесноту связи в условиях **линейной зависимости** между изучаемыми признаками. При наличии **нелинейной связи** он может привести к неверным выводам о степени тесноты связи (его значение занижено) .

Парный регрессионный анализ

- Сутью регрессионного анализа является описание "технологии" влияния признаков-факторов на признак-результат, который в конкретных практических задачах выступает объектом управления.
- Регрессионный анализ предполагает теоретический анализ природы изучаемого явления с целью определения круга факторов, оказывающих влияние на поведение результативного признака. На базе корреляционного анализа выявляется наличие статистически значимых связей в конкретных условиях места и времени. Затем строится уравнение регрессии (аналитическая форма изучаемой зависимости), которое при определенных условиях может быть признано статистической моделью связи между признаками.

Корреляционный и регрессионный анализ данных.
Множественный регрессионный анализ.

- Уравнение регрессии – это математическая функция, описывающая зависимость условного среднего значения результативной (зависимой) переменной от заданных значений факторных (независимых) переменных. Таким образом, уравнение регрессии отражает основную тенденцию связи, характерную для изучаемой статистической совокупности в целом.

В регрессионном анализе можно выделить три составляющие:

- - определение типа функции (структуры модели) для описания изучаемой зависимости;
- - расчет неизвестных параметров уравнения регрессии;
- - оценку качества модели.

До широкого распространения компьютерных технологий перечисленные элементы являлись последовательными этапами анализа. В современных условиях все процедуры выполняются комплексно.

Первый этап регрессионного анализа – поиск линии регрессии, которая бы лучшим образом аппроксимировала поле корреляции.

Необходимо учитывать природу изучаемых показателей, специфику их взаимосвязи, свойства математических функций.

Современные ППП позволяют одновременно построить несколько видов уравнений, а затем, пользуясь специальными критериями, отобрать лучшую модель.

В качестве критерия могут быть использованы:

максимальное значение коэффициента детерминации, максимальное значение F-критерия Фишера, минимальное значение остаточной дисперсии, минимальное значение стандартной ошибки уравнения, минимальное значение средней ошибки аппроксимации.

Для аналитического описания связи между признаками могут быть использованы следующие виды уравнений:

- 1) $\bar{y} = a_0 + a_1 x$ – прямая, линейная функция;
- 2) $\bar{y} = a_0 + a_1 x^2 + a_2 x$ – парабола;
- 3) $\bar{y} = a_0 + a_1 \frac{1}{x}$ – гипербола;
- 4) $\bar{y} = a_0 x^{a_1}$ – степенная функция;
- 5) $\bar{y} = \exp(a_0 + a_1 x)$ – экспонента и др.

Некоторые задачи корреляционно-регрессионного анализа, а также возможности ППП, делают необходимым выполнение операции линеаризации уравнений, т.е. приведение их к линейному виду путем логарифмирования. Производится замена признака-фактора и признака-результата их натуральными логарифмами. При проведении анализа с использованием линеаризации необходимо помнить о том, что все показатели и графические изображения рассчитываются и строятся для логарифмов признаков.

Простейшим видом уравнения регрессии является парная линейная регрессия

$$\bar{y} = a_0 + a_1 x + \varepsilon$$,

где $\bar{y}$ – расчетное, теоретическое значение признака-результата;

a_0, a_1 – параметры уравнения регрессии;

ε – случайная величина.

Присутствие в уравнении ε связано с рядом причин, среди которых: наличие признаков-факторов, не включенных в данное уравнение; неправильное описание структуры модели; ошибки измерений и др.