**Исследование посещаемости WEB сайта**

**Теоретическая часть работы**

**Основные задачи корреляционно-регрессионного анализа**

Все явления и процессы, характеризующие социально-экономическое развитие и составляющие единую систему национальных счетов, тесно взаимосвязаны и взаимозависимы между собой.

В статистике показатели, характеризующие эти явления, могут быть связаны либо корреляционной зависимостью, либо быть независимыми Корреляционная зависимость является частным случаем стохастической зависимости, при которой изменение значений факторных признаков (х 1 х2 ..., хn ) влечет за собой изменение среднего значения результативного признака.

Корреляционная зависимость исследуется с помощью методов корреляционного и регрессионного анализов.

Корреляционный анализ изучает взаимосвязи показателей и позволяет решить следующие задачи.

1. Оценка тесноты связи между показателями с помощью парных, частных и множественных коэффициентов корреляции

**2. Оценка уравнения регрессии.**

Основной предпосылкой применения корреляционного анализа является необходимость подчинения совокупности значений всех факторных (х1 х2 .... хn) и результативного (У) признаков r-мерному нормальному закону распределения или близость к нему. Если объем исследуемой совокупности достаточно большой ( n > 50), то нормальность распределения может быть подтверждена на основе расчета и анализа критериев Пирсона, Боярского, Колмогорова, чисел Вастергарда и т. д. Если n < 50, то закон распределения исходных данных определяется на базе построения и визуального анализа поля корреляции. При этом если в расположении точек имеет место линейная тенденция, то можно предположить, что совокупность исходных данных подчиняется нормальному распределению.

Целью регрессионного анализа является оценка функциональной зависимости условного среднего значения результативного признака (У) от факторных (х1. Х2..., хn).

Основной предпосылкой регрессионного анализа является то, что только результативный признак (У) подчиняется нормальному закону распределения, а факторные признаки х1. Х2..., хn могут иметь произвольный закон распределения. В анализе динамических рядов в качестве факторного признака выступает время t При этом в регрессионном анализе заранее подразумевается наличие причинно-следственных связей между результативным (У) и факторными х1. Х2..., хn признаками.

Уравнение регрессии, или статистическая модель связи социально-экономических явлений, выражаемая функцией Y=f(х1. Х2..., хn) является достаточно адекватным реальному моделируемому явлению или процессу в случае соблюдения следующих требований их построения.

1. Совокупность исследуемых исходных данных должна быть однородной и математически описываться непрерывными функциями.

2. Возможность описания моделируемого явления одним или несколькими уравнениями причинно-следственных связей.

3. Все факторные признаки должны иметь количественное (цифровое) выражение.

4. Наличие достаточно большого объема исследуемой выборочной совокупности.

5. Причинно-следственные связи между явлениями и процессами следует описывать линейной или приводимой к линейной формой зависимости.

6. Отсутствие количественных ограничений на параметры модели связи.

7. Постоянство территориальной и временной структуры изучаемой совокупности.

Соблюдение данных требований позволяет исследователю построить статистическую модель связи, наилучшим образом аппроксимирующую моделируемые социально-экономические явления и процессы.

**Корреляция случайных величин**

Прямое токование термина корреляция — стохастическая, вероятная, возможная связь между двумя (парная) или несколькими (множественная) случайными величинами.

Для числовой оценки возможной связи между двумя случайными величинами: Y(со средним My и среднеквадратичным отклонением Sy) и — X (со средним Mx и среднеквадратичным отклонением Sx) принято использовать так называемый коэффициент корреляции

Rxy= .



Этот коэффициент может принимать значения от -1 до +1 — в зависимости от тесноты связи между данными случайными величинами.

Если коэффициент корреляции равен нулю, то X и Y называют некоррелированными. Считать их независимыми обычно нет оснований — оказывается, что существуют такие, как правило — нелинейные связи величин, при которых Rxy = 0, хотя величины зависят друг от друга. Обратное всегда верно — если величины независимы, то Rxy = 0. Но, если модуль Rxy = 1, то есть все основания предполагать наличие линейной связи между Y и X. Именно поэтому часто говорят о линейной корреляции при использовании такого способа оценки связи между СВ.

В отдельных случаях приходится решать вопрос о связях нескольких (более 2) случайных величин или вопрос о множественной корреляции.

Пусть X, Y и Z - случайные величины, по наблюдениям над которыми мы установили их средние Mx, My,Mz и среднеквадратичные отклонения Sx, Sy, Sz.

Тогда можно найти парные коэффициенты корреляции Rxy, Rxz, Ryz по приведенной выше формуле. Но этого явно недостаточно - ведь мы на каждом из трех этапов попросту забывали о наличии третьей случайной величины! Поэтому в случаях множественного корреляционного анализа иногда требуется отыскивать т. н. частные коэффициенты корреляции — например, оценка виляния Z на связь между X и Y производится с помощью коэффициента

Rxy.z =



И, наконец, можно поставить вопрос — а какова связь между данной СВ и совокупностью остальных? Ответ на такие вопросы дают коэффициенты множественной корреляции Rx.yz, Ry.zx, Rz.xy, формулы для вычисления которых построены по тем же принципам — учету связи одной из величин со всеми остальными в совокупности.

На сложности вычислений всех описанных показателей корреляционных связей можно не обращать особого внимания - программы для их расчета достаточно просты и имеются в готовом виде во многих ППП современных компьютеров. Например программное обеспечение «Олимп» с помощью которого производится ряд расчетов в этой работе.

**Линейная регрессия**

В тех случаях, когда из природы процессов в модели или из данных наблюдений над ней следует вывод о нормальном законе распределения двух СВ - Y и X, из которых одна является независимой, т. е. Y является функцией X, то возникает соблазн определить такую зависимость “формульно”, аналитически.

В случае успеха нам будет намного проще вести моделирование. Конечно, наиболее заманчивой является перспектива линейной зависимости типа Y = a + b•X .

Подобная задача носит название задачи регрессионного анализа и предполагает следующий способ решения.

Выдвигается следующая гипотеза:

H0: случайная величина Y при фиксированном значении величины X распределена нормально с математическим ожиданием

My = a + b•X и дисперсией Dy, не зависящей от X.

При наличии результатов наблюдений над парами Xi и Yi предварительно вычисляются средние значения My и Mx, а затем производится оценка коэффициента b в виде

b = = Rxy



что следует из определения коэффициента корреляции. После этого вычисляется оценка для a в виде {2 - 16}

и производится проверка значимости полученных результатов. Таким образом, регрессионный анализ является мощным, хотя и далеко не всегда допустимым расширением корреляционного анализа, решая всё ту же задачу оценки связей в сложной системе.

Теперь более подробно рассмотрим множественную или многофакторную регрессию. Нас интересует только линейная модель вида: Y=A0+A1X1+A2X2+…..AkXk.

Изучение связи между тремя и более связанными между собой признаками носит название множественной (многофакторной) регрессии. При исследовании зависимостей методами множественной регрессии задача формулируется так же, как и при использовании парной регрессии, т. е. требуется определить аналитическое выражение связи между результативным признаком (У) и факторными признаками (х1 х2, х3 ..., хn) найти функцию: Y=f(х1. Х2..., хn)

Построение моделей множественной регрессии включает несколько этапов:

• выбор формы связи (уравнения регрессии):

• отбор факторных признаков:

• обеспечение достаточного объема совокупности для получения несмещенных оценок.

Рассмотрим подробнее каждый из них.

Выбор формы связи затрудняется тем, что, используя математический аппарат, теоретически зависимость между признаками может быть выражена большим числом различных функций.

Выбор типа уравнения осложнен тем, что для любой формы зависимости выбирается целый ряд уравнений, которые в определенной степени будут описывать эти связи. Некоторые предпосылки для выбора определенного уравнения регрессии получают на основе анализа предшествующих аналогичных исследований или на базе анализа подобных работ в смежных отраслях знаний. Поскольку уравнение регрессии строится главным образом для объяснения и количественного выражения взаимосвязей, оно должно хорошо отражать сложившиеся между исследуемыми факторами фактические связи,

Наиболее приемлемым способом определения вида исходного уравнения регрессии является метод перебора различных уравнений.

Сущность данного метода заключается в том, что большое число уравнений (моделей) регрессии, отобранных для описания связей какого-либо социально-экономического явления или процесса, реализуется на ЭВМ с помощью специально разработанного алгоритма перебора с последующей статистической проверкой, главным образом на основе t-крнтерия Стьюдeнта и F-критерия Фишера. Способ перебора является достаточно трудоемким и связан с большим объемом вычислительных работ. Практика построения многофакторных моделей взаимосвязи показывает, что все реально существующие зависимости между социально-экономическими явлениями можно описать, используя пять типов моделей:

линейная: Y=A0+A1X1+….AkXk

степенная

показательная

параболическая

гиперболическая

Основное значение имеют линейные модели в силу простоты и логичности их экономической интерпретации. Нелинейные формы зависимости приводятся к линейным путем линеаризации.

Важным этапом построения уже выбранного уравнения множественной регрессии являются отбор и последующее включение факторных признаков. Сложность формирования уравнения множественной регрессии заключается в том, что почти все факторные признаки находятся в зависимости один от другого. Проблема размерности модели связи, т. е. определение оптимального числа факторных признаков, является одной из основных проблем построения множественного уравнения регрессии. С одной стороны, чем больше факторных признаков включено в уравнение, тем оно лучше описывает явление. Однако модель размерностью 100 и более факторных признаков сложно реализуема и требует больших затрат машинного времени. Сокращение размерности модели за счет исключения второстепенных, экономически и статистически несущественных факторов способствует простоте и качеству ее реализации. В то же время построение модели регрессии малой размерности может привести к тому, что такая модель будет недостаточно адекватна исследуемым явлениям и процессам. Проблема отбора факторных признаков для построения моделей взаимосвязи может быть решена на основе эвристических или многомерных статистических методов анализа.

Метод экспертных оценок как эвристический метод анализа основных макроэкономических показателей, формирующих единую междуна- , родную систему расчетов, основан на интуитивно-логических предпосылках, содержательно-качественном анализе. Анализ экспертной информации проводится на базе расчета и анализа непараметрических показателей связи: ранговых коэффициентов корреляции Спирмена, Кендалла и конкордации .

Наиболее приемлемым способом отбора факторных признаков является шаговая регрессия (шаговый регрессионный анализ). Сущность метода шаговой регрессии заключается в последовательном включении факторов в уравнение регрессии и последующей проверке их значимости. Факторы поочередно вводятся в уравнение так называемым "прямым методом". При проверке значимости введенного фактора определяется, насколько уменьшается сумма квадратов остатков и увеличивается величина множественного коэффициента корреляции . одновременно используется и обратный метод, т.е. , исключение факторов, ставших незначимыми на основе t-критерия Стьюдента. Фактор является незначимым, если его включение в уравнение регрессии только изменяет значение коэффициентов регрессии, не уменьшая суммы квадратов остатков и не увеличивая их значения. Если при включении в модель соответствующего факторного признака величина множественного коэффициента корреляции увеличивается, а коэффициент регрессии не изменяется (или меняется несущественно), то данный признак существен и его включение в уравнение регрессии необходимо.

Если же при включении в модель факторного признака коэффициенты регрессии меняют не только величину, но и знаки, а множественный

коэффициент корреляции не возрастает, то данный факторный признак признается нецелесообразным для включения в модель связи.

Сложность и взаимное переплетение отдельных факторов, обусловливающих исследуемое экономическое явление (процесс), могут проявляться в так называемой мультиколлинеарности. Под

мультиколлинеарностью понимается тесная зависимость между факторными признаками, включенными в модель.

Наличие мультиколлинеарности между признаками приводит к:

• искажению величины параметров модели, которые имеют тенденцию к завышению;

• изменению смысла экономической интерпретации коэффициентов регрессии;

. слабой обусловленности системы нормальных уравнений;

. осложнению процесса определения наиболее существенных факторных признаков.

Одним из индикаторов определения наличия мультиколлинеарности между признаками является превышение парным коэффициентом корреляции величины 0,8 .

Устранение мультиколлинеарности может реализовываться через исключение из корреляционной модели одного или нескольких линейно-связанных факторных признаков или преобразование исходных факторных признаков в новые, укрупненные факторы.

Вопрос о том, какой из факторов следует отбросить, решается на основании качественного и логического анализов изучаемого явления.

Качество уравнения регрессии зависит от степени достоверности и надежности исходных данных и объема совокупности. Исследователь должен стремиться к увеличению числа наблюдений, так как большой объем наблюдений является одной из предпосылок построения адекватных статистических моделей.

Аналитическая форма выражения связи результативного признака и ряда факторных называется многофакторным (множественным) уравнением регрессии, или моделью связи.

Уравнение линейной множественной регрессии имеет вид:

Y=A0+A1X1+….AkXk

Коэффициенты Аn вычисляются при помощи систем нормальных уравнений. Например система нормальных уравнений для вычисления коэффициентов регрессии для уравнения линейной регрессии с двумя факторными признаками:



где An=an

Общий вид нормальных уравнений для расчета коэффициентов регрессии:



Оценка существенности связи, принятие решения на основе уравнения регрессии.

Проверка адекватности моделей, построенных на основе уравнений регрессии, начинается с проверки значимости каждого коэффициента регрессии.

Значимость коэффициентов регрессии осуществляется с помощью

t-критерия Стьюдента:



- дисперсия коэффициента регрессии.

Параметр модели признается статистически значимым, если tp>tкр

Наиболее сложным в этом выражении является определение дисперсии, которая может быть рассчитана двояким способом.

Наиболее простой способ, выработанный методикой экспериментирования, заключается в том, что величина дисперсии коэффициента регрессии может быть приближенно определена по выражению:



- дисперсия результативного признака:

k - число факторных признаков в уравнении.

Наиболее сложным этапом, завершающим регрессионный анализ, является интерпретация уравнения, т. е. перевод его с языка статистики и математики на язык экономиста.

Интерпретация моделей регрессии осуществляется методами той отрасли знаний, к которой относятся исследуемые явления. Но всякая интерпретация начинается со статистической оценки уравнения регрессии в целом и оценки значимости входящих в модель факторных признаков, т. е. с выяснения, как они влияют на величину результативного признака. Чем больше величина коэффициента регрессии, тем значительнее влияние данного признака на моделируемый. Особое значение при этом имеет знак перед коэффициентом регрессии. Знаки коэффициентов регрессии говорят о характере влияния на результативный признак. Если факторный признак имеет знак плюс, то с увеличением данного фактора результативный признак возрастает; если факторный признак со знаком минус, то с его увеличением результативный признак уменьшается. Интерпретация этих знаков полностью определяется социально-экономическим содержанием моделируемого (результативного) признака. Если его величина изменяется в сторону увеличения, то плюсовые знаки факторных признаков имеют положительное влияние. При изменении результативного призна-л-1 в сторону снижения положительное значение имеют минусовые знаки факторных признаков. Если экономическая теория подсказывает, что факторный признак должен иметь положительное значение, а он со знаком минус, то необходимо проверить расчеты параметров уравнения регрессии. Такое явление чаще всего бывает в силу допущенных ошибок при решении. Однако следует иметь в виду, что при анализе совокупного влияния факторов, при наличии взаимосвязей между ними характер их влияния может меняться. Для того чтобы быть уверенным, что факторный признак изменил знак влияния, необходима тщательная проверка решения данной модели, так как часто знаки могут меняться в силу допустимых ошибок при сборе или обработке информации.

При адекватности уравнения регрессии исследуемому процессу возможны следующие варианты.

1. Построенная модель на основе ее проверки по F-критерию Фишера в целом адекватна, и все коэффициенты регрессии значимы. Такая модель может быть использована для принятия решений к осуществлению прогнозов.

2. Модель по F-критерию Фишера адекватна, но часть коэффициентов регрессии незначима. В этом случае модель пригодна для принятия некоторых решений, но не для производства прогнозов.

3. Модель по F-критерию Фишера адекватна, но все коэффициенты регрессии незначимы. Поэтому модель полностью считается неадекватной. на ее основе не принимаются решения и не осуществляются прогнозы.

**Практическая часть работы**

**1. Описание объекта**

В нашем случае объектом исследования является совокупность наблюдений за посещаемостью WEB сайта Комитета по делам семъи и молодежи Правительства г. Москвы www.telekurs.ru/ismm. Тематика сайта – это предоставление социально незащищенным слоям населения: молодежи, студентам информации о трудоустройстве в Москве. Информация ежедневно обновляется, приблизительно 200 новых вакансий в день. Также на сайте содержится информация о текущих программах правительства г. Москвы направленных на поддержку указанных выше категорий населения. Моделируемым показателем является N- количество человек в день посетивших сайт.

**2. Факторы формирующие моделируемое явление**

Отбор факторов для модели осуществляется в два этапа. На первом идет анализ, по результатам которого исследователь делает вывод о необходимости рассмотрения тех или иных явлений в качестве переменных, определяющих закономерности развития исследуемого процесса, на втором – состав предварительно отобранных факторов уточняется непосредственно по результатам статистического анализа.

Полученные данные с помощью программы наблюдения за компьютерной сетью (Net Medic, Net lab) являются не совсем точными, но довольно близки к реальным и по этому будем считать, что они дают представление о характере процесса. (получение более точных данных было для автора невозможно в связи с недостаточной технической базой) Из совокупности этих факторов я отобрал следующие :

Зависимый фактор:

N- количество человек в день посетивших сайт.

Для модели в абсолютных показателях

Независимые факторы:

P - Загруженность внутренней сети (чел/день)

S – Cкорость обмена данными в сети Кбит/сек

V – Кол-во вакансий на текущий день

B – Количество «Баннеров» – рекламных ссылок на исследуемый сайт.

Данные представлены в таблице 1.

Таблица 1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| № Объекта  наблюдения | N  Кол-во человек в день | P  Загруженность внутренней сети  (чел/ден) | S  Скорость обмена данными в сети Кбит/сек | V  Кол-во вакансий на текущий день. | B  Кол-во баннеров |
| 1 | 11 | 651 | 2627 | 165 | 4 |
| 2 | 18 | 1046 | 3045 | 400 | 4 |
| 3 | 19 | 944 | 2554 | 312 | 5 |
| 4 | 11 | 1084 | 4089 | 341 | 4 |
| 5 | 15 | 1260 | 6417 | 496 | 7 |
| 6 | 10 | 1212 | 4845 | 264 | 8 |
| 7 | 12 | 254 | 923 | 78 | 1 |
| 8 | 14 | 1795 | 9602 | 599 | 13 |
| 9 | 9 | 2851 | 12542 | 622 | 12 |
| 10 | 15 | 1156 | 6718 | 461 | 9 |

**3. Анализ матрицы коэффициентов парных корреляций**

Таблица 2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| № фактора | N | P | S | V | B |
| N | 1.00 | -0.22 | -0.06 | 0.44 | 0.12 |
| P | -0.22 | 1.00 | 0.91 | 0.68 | 0.74 |
| S | -0.06 | 0.91 | 1.00 | 0.86 | 0.91 |
| V | 0.44 | 0.68 | 0.86 | 1.00 | 0.85 |
| B | 0.12 | 0.74 | 0.91 | 0.85 | 1.00 |

Из таблицы 2 находим тесно коррелирующие факторы. Налицо мультиколлениарность факторов P и S ( 0.91 ). Оставим только один фактор P . И действительно если скорость в сети высокая то она может без значительных задержек во времени обработать значительное кол-во запросов от пользователей, значит чем больше скорость в сети тем больше в ней пользователей. Тем загруженее сеть.

**4. Построение уравнения регрессии**

Используя программное обеспечение «ОЛИМП» (которое в свою очередь использует для расчетов указанные выше принципы и формулы чем значительно облегчает нам жизнь), найдем искомое уравнение множественной регрессии, исключив из расчетов, как указывалось выше, факторы S – скорость сети (чел/день )

Путем перебора возможных комбинаций оставшихся факторных признаков получим следующую модель:

Функция N = +12.567-0.005\*P+0.018\*V

Оценки коэффициентов линейной регрессии

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № | Значение | Дисперсия | Среднеквадратическое отклонение | Значение tрасч |
| 1 | 12.57 | 2.54 | 1.59 | 7.88 |
| 2 | -0.01 | 0 | 0 | -3.60 |
| 3 | 0.02 | 0 | 0 | 4.07 |

Кpитические значения t-pаспpеделения пpи 8 степенях свободы имеют следующие значения:

веpоятность t-значение

0.900 1.400

0.950 1.863

0.990 2.887

В нашей модели |tрасч |> tкритич у всех коэффициентов регрессии значит можно утверждать, что модель является адекватной моделируемому явлению, т.е. гипотеза о значимости уравнения не отвергается, о чем говорят также данные выдаваемые компьютером:

Характеристики остатков

Среднее значение...................………….. -0.000

Оценка дисперсии...................…………. 3.6

Оценка приведенной дисперсии......…. 4.95

Средний модуль остатков...........……… 1.391

Относительная ошибка аппроксимации. 9.898

Критерий Дарбина-Уотсона...........……. 1.536

Коэффициент детерминации...........…… 0.690

F - значение ( n1 = 3, n2 = 8).………. 143

Гипотеза о значимости уравнения не отвергается с вероятностью 0.950

**5. Смысл модели**

При увеличении количества вакансий в день, количество посетивших сайт людей будет увеличиваться . Это означает что в настоящий момент сайт не полностью удовлетворяет запросы пользователей, что необходимо увеличить количество вакансий, но в связи со сложившимся в экономике России положением это представляется проблематичным.

При увеличении загруженности внутренней сети в которой расположен сервер содержащий исследуемый сайт количество людей посетивших сайт будет уменьшатся из-за снижения скорости доступа к нему а также из-за возможных перегрузках в узлах сети, в связи с чем сервер содержащий сайт может не отвечать на запросы пользователей. Также с перегрузкой связаны различные сбои в работе системы, что отрицательно сказывается на работе сайта. Коэффициент детерминации у линейной модели - 0.69. Это означает , что факторы , вошедшие в модель объясняют изменение количества посетивших сайт людей на 69%. Следовательно значения полученные с помощью линейной модели близки к фактическим.

**Список литературы**

«Теория статистики» учебник под редакцией проф. Р.А.Шмойловой Издательство «Финансы и статистика» 1996 г.