СОДЕРЖАНИЕ

Содержание

Введение

1. Теоретическая часть
   1. Теоретические основы прикладного регрессионного анализа
   2. Проверка предпосылок и предположений регрессионного анализа
      1. Проверка случайности
      2. Проверка стационарности
   3. Обнаружение выбросов в выборке
   4. Мультиколлинеарность переменных
      1. Рекомендации по устранению мультиколлинеарности
      2. Доверительные интервалы для уравнения регрессии
      3. Определение доверительного интервала для истинного значения уравнения регрессии
      4. Свойства доверительных интервалов
   5. Адекватность модели
2. Практическая часть

Вывод

Список литературы

ВВЕДЕНИЕ

Общее назначение множественной регрессии (этот термин был впервые использован в работе Пирсона - Pearson, 1908) состоит в анализе связи между несколькими независимыми переменными (называемыми также регрессорами или предикторами) и зависимой переменной. Например, агент по продаже недвижимости мог бы вносить в каждый элемент реестра размер дома (в квадратных футах), число спален, средний доход населения в этом районе в соответствии с данными переписи и субъективную оценку привлекательности дома. Как только эта информация собрана для различных домов, было бы интересно посмотреть, связаны ли и каким образом эти характеристики дома с ценой, по которой он был продан. Например, могло бы оказаться, что число спальных комнат является лучшим предсказывающим фактором (предиктором) для цены продажи дома в некотором специфическом районе, чем "привлекательность" дома (субъективная оценка). Могли бы также обнаружиться и "выбросы", т.е. дома, которые могли бы быть проданы дороже, учитывая их расположение и характеристики.

Специалисты по кадрам обычно используют процедуры множественной регрессии для определения вознаграждения адекватного выполненной работе.

Как только эта так называемая линия регрессии определена, аналитик оказывается в состоянии построить график ожидаемой (предсказанной) оплаты труда и реальных обязательств компании по выплате жалования. Таким образом, аналитик может определить, какие позиции недооценены (лежат ниже линии регрессии), какие оплачиваются слишком высоко (лежат выше линии регрессии), а какие оплачены адекватно.

1. ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

1.1 Теоретические основы прикладного регрессионного анализа

Регрессионный анализ применяется для построения математических зависимостей объектов, явлений по результатам экспериментальных данных, полученных на основе проведения активного или пассивного экспериментов.

Предполагается, что математическая зависимость относится к определенному классу функций с несколькими неизвестными параметрами. В общем виде эти функции представим в виде:

,



где - вектор зависимой (выходной) переменной размерностью ;



- матрица независимых (входных) переменных размерностью ;



- вектор неизвестных параметров размерностью ;



- вектор возмущений размерностью ;



- количество независимых переменных;



- количество экспериментальных данных;



- класс функциональных зависимостей.



В зависимости – является случайной величиной, значения могут рассматриваться либо как фиксированные, либо как случайные. При этом ожидаемое значение одной случайной переменной соотносится с наблюдаемыми значениями других случайных переменных в виде условной регрессии.



Рассмотрим зависимость между случайными величинами и , представленную в виде некоторой таблицы наблюдений значений и .



Перенося табличные значения и на плоскость , получаем поле корреляции, приведенное на рисунке 3.1



Рисунок 1.1 — Экспериментальное уравнение регрессии

Разобьем диапазон изменения на -равных интервалах . Все точки, попавшие в интервал , отнесем к середине интервала , в результате получаем трансформированное поле корреляции.



Определим частичные средние арифметические для каждого значения :



,



где - число точек, оказавшихся в интервале, причем , где



- общее число наблюдений.



Соединим последовательно точки с координатами и отрезками прямых. Полученная ломаная линия называется эмпирической линией регрессии по ; она показывает, как в среднем меняется с изменением . Предельное положение эмпирической линии регрессии, к которому она стремится при неограниченном увеличении числа наблюдений и одновременном уменьшении , называется предельной теоретической линией регрессии. Ее нахождение и составляет основную задачу регрессионного анализа. Отметим, что по линии регрессии невозможно точно определить значение по в одном опыте. Однако зависимость позволяет определить в среднем значение при многократном повторении опыта при фиксированном значении . В регрессионном анализе рассматривается связь между одной переменной, называемой зависимой, и несколькими другими, называемыми независимыми. Эта связь представляется в виде математической модели, т.е. в виде функции регрессии. Если функция линейна относительно параметров, но не обязательно линейна относительно независимых переменных, то говорят о линейной модели. В противном случае нелинейная. Статистическими проблемами обработки в регрессионном анализе являются:



а) Получение наилучших точечных и интервальных оценок неизвестных параметров регрессионного анализа;

б) Проверка гипотез относительно этих параметров;

в) Проверка адекватности;

г) Проверка множества предполагаемых предположений.

Исследуемый объект представлен на рисунке 3.2



Рисунок 1.2 — Вид исследуемого объекта

Для корректного использования регрессионного анализа существует следующие предпосылки и следующие допущения на свойства регрессионной ошибки , ; - значение зависимой переменной, полученное подстановкой в уравнение , , ; - количество экспериментальных данных, - количество независимых переменных:



Приведем свойства и предпосылки регрессионной ошибки:

а) Свойства регрессионной ошибки:

1) В каждом опыте имеет нормальный закон распределения;



, .



2) В каждом опыте математическое ожидание равно нулю;



, .



3) Во всех опытах дисперсия постоянна и одинакова;



, .



4) Во всех опытах ошибки независимы.



, .



б) предпосылки регрессионной ошибки:

1). Матрица наблюдений имеет полный ранг;



.



2). Структура модели адекватна истинной зависимости;

3). Значения случайной ошибки не зависят от значений регрессоров ;



4). Ошибки регистрации регрессоров пренебрежимо малы по сравнению со случайной ошибкой .



1.2 Проверка предпосылок и предположений регрессионного анализа

Регрессионный анализ является одним из самых распространённых методов обработки результатов наблюдений. Он служит основой для целого ряда разделов математической статистики и методов обработки данных. Регрессионный анализ базируется на ряде предположений и предпосылок, нарушение которых приводит к некорректному его использованию и ошибочной интерпретации результатов.

Если F-критерий и показал, что подгонка модели в целом является удовлетворительной; целесообразно провести анализ остатков для проверки соблюдений предпосылок и предположений.

В этом случае исследуется набор отклонений между экспериментальными и предсказанными значениями зависимой переменной,

.



Проверка предпосылок и предположений регрессионного анализа включает в себя следующие задачи:

1) оценка случайности зависимой переменной;

2) оценка стационарности и эргодичности зависимых и независимых переменных;

3) Проверка гипотезы о нормальности распределения ошибок *E*;

4) Обнаружение выбросов;

5) Проверка постоянства математического ожидания и дисперсии ошибок;

6) Оценка коррелированности остатков;

7) Обнаружение мультиколлинеарности.

1.2.1 Проверка случайности

Построение моделей методом множественного регрессионного анализа требуется выполнение предположения случайности и в нормальной линейной модели вида



где – вектор наблюдений зависимой переменной;



– матрица наблюдений независимых переменных;



– вектор неизвестных коэффициентов;



– вектор ошибок.



Задача проверки случайности может быть разбита на 2 подзадачи:

1) проверка случайности собственной величины Y;

2) проверка случайности выборки, то есть допущения об отсутствии существенного смещения средней величины во времени.

Первая подзадача решается с использованием критерия серий. Для этой цели последовательность наблюдений величины *Y* представляют последовательностью нулей и единиц, где единицей обозначают значение, превышающее среднее или медиану, и нулем, собственно, значение меньшее медианы. После обозначения вектор наблюдений преобразуется в последовательность серий где – количество подряд идущих элементов одного вида, *i* – номер серии.



Доказано, что при распределение величины *r* близится к нормальному с характеристиками



Тогда с вероятностью 0,954 теоретическое число серий *r* будет находиться в пределах



Если фактическое значение попадает в указанные пределы, то *Y* можно считать случайной величиной.



Серией называется последовательность наблюдаемых значений, перед которыми и после которых расположены наблюдаемые значения другой категории. Если последовательность *N* наблюдений представляет собой независимые наблюденные значения одной и той же случайной величины, т.е. вероятность знаков (+) и (–) не меняется от одного наблюдения к другому, то выборочное распределение числа серий в последовательности есть случайная величина *r* со средним значением

(3.1)



и дисперсией

(3.2)



Здесь – число наблюдений со знаком (+), – число наблюдений со знаком (–).



Когда соотношения (3.1) и (3.2) принимают вид



Для решения второй подзадачи используется метод последовательных разностей. Элементы исследуемой выборки располагаются в порядке получения наблюдений и для них вычисляются выборочные среднее и дисперсия



Определяют разности между соседними наблюдениями



и математическое ожидание квадрата разности



где – оценка генеральной дисперсии.



Фактическая величина критерия случайности выборки

.



Теоретическое значение критерия



При для конкретного *N* гипотеза случайности отвергается.



1.2.2 Проверка стационарности

Анализ случайных процессов может производиться осреднением величин по ансамблю выборочных реализаций или по одной реализации.

Поскольку на практике проверка по ансамблю достаточно длинных выборочных реализаций неосуществима, то для использования тестов проверки стационарности процесса принимается ряд допущений:

а) проверка заключается в исследовании поведения не ансамбля, а его отдельных реализаций; это означает, что доказательство внутренней стационарности отдельных реализаций может служить доказательством стационарности случайного процесса, которому принадлежит эта реализация;

б) для большинства процессов достаточно проверить слабую стационарность, поскольку, во-первых, для эффективного использования спектрального и корреляционного анализа случайных процессов достаточно выполнения условия слабой стационарности, а во-вторых, для реальных процессов обычно слабая стационарность влечет за собой и строгую; если процесс определяется нормальной плотностью, то это доказательство осуществляется автоматически, поскольку все моменты высших порядков полностью определяются средним и автокорреляционной функцией;

в) на практике часто стационарность автокорреляционной функции обеспечивается стационарностью дисперсии.

Учитывая эти допущения, проверку стационарности осуществляют исследованием одной реализации .



Для этого реализация делится на *N* равных интервалов таких, что её участки в пределах каждого интервала можно считать независимыми. Для всех интервалов вычисляются средние значения и средние значения квадратов, из которых составляются две последовательности и затем их проверяют на наличие тренда.



Если известно выборочное распределение, то для проверки можно использовать существующие непараметрические критерии (t-критерий Стьюдента, -критерий Пирсона, F-критерий Фишера), однако в обычной ситуации проверка стационарности осуществляется при высокой неопределенности относительно исследуемого процесса. В этом случае целесообразно использовать непараметрические критерии, например, критерий серий и критерий тренда



Критерий тренда основан на подсчете числа случаев, когда для в последовательности *N* наблюденных значений величины *x*.



Такое неравенство называется инверсией, а их число *k* определяется из соотношения

,



где



Число инверсий есть также случайная величина со средним



и дисперсией

.



Область принятия гипотезы ограничена интервалом .



Критерий тренда обладает большей мощностью при выявлении монотонного тренда, однако при выявлении колебательного тренда его мощность невелика, в этом случае целесообразнее использовать критерий серий.

Критерии проверки гипотезы стационарности обладают рядом особенностей:

1) Нет необходимости знать ширину полосы частот исследуемых процессов;

2) Не требуется точно знать время осреднения, использованное для вычисления средних и квадратов отклонений от средних;

3) Для проверки не обязательно, чтобы исследуемые процессы были полностью случайными. При изучении процессов может возникнуть случай, когда независимость от времени средних и квадратов не является достаточным условием для утверждения о независимости от времени автокорреляционной функции.

1.3 Обнаружение выбросов в выборке

Выбросом среди остатков представляет собой остаток, который значительно превосходит по абсолютной величине остальные и отличается от среднего по остаткам на три, четыре или даже более стандартных отклонений.

Для обнаружения выбросов необходимо построить график остатков, определённых по формуле



В случае если , данная точка будет характеризовать выброс. Следует отметить, что иногда выброс может дать полезную информацию. В этом случае необходимо более тщательное исследование выбросов, а не механическое их отбрасывание. Выбросы должны быть исключены сразу если выясняется, что они вызваны такими причинами, как ошибки в регистрации данных, неудовлетворительная настройка аппаратуры и т.д. Если имеется не одно аномальное измерение, то критерий их не обнаруживает, особенно если анализируется менее 30 измерений.



1.4 Мультиколлинеарность переменных

Одно из основных предположений регрессионного анализа относится к матрице исходных данных: среди независимых переменных не должно быть линейно зависимых. Это требование необходимо для вычисления оценки методом наименьших квадратов.



Мультиколлинеарность приводит к:

1)снижению точности, дисперсия оценок увеличивается, параметры модели коррелированны, что приводит к трудностям в интерпретации модели;

2)оценки коэффициентов становятся чувствительны к особенностям множества выборочных данных.

Причиной мультиколлинеарности могут служить:

1) наличие автокорреляции в ряду наблюдений;

2) корреляция между переменными;

3) высокий уровень помех.

Под мультиколлинеарностью будем понимать сопряженность независимых переменных, это означает "почти линейную зависимость" векторов , т.е. существование чисел таких, что:



(3.3)



Когда равенство (3.3) имеет место, говорят о строгой мультиколлинеарности.

При наличии мультиколлинеарности оценки МНК становятся положительными, т.е. дисперсия оценок будет весьма большой. При наличии (3.3) матрица становится плохо обусловленной, в частности , т.е. .



1.4.1 Рекомендации по устранению мультиколлинеарности

Наиболее простой способ устранения мультиколлинеарности – исключение одной переменной из пары переменных, коэффициент корреляции между которыми больше 0,8.

Простейшие рекомендации по устранению мультиколлинеарности сводятся к сокращению рассматриваемого множества объясняющих переменных за счет тех из них, которые линейно связаны с уже включенными в модель. Выполнение этих рекомендаций ведет к построению сокращенной модели, которая не всегда соответствует требованию наблюдательности и управляемости. Чтобы избежать нежелательных эффектов мультиколлинеарности, сохранив при этом весь интересующий нас набор объясняющих переменных, предлагается увеличить размеры выборки путем получения дополнительной информации. Ясно, что не любое произвольное увеличение выборки ведет к ослаблению эффектов мультиколлинеарности.

Часто для устранения мультиколлинеарности используют приемы, основанные на предварительном преобразовании исходных данных путем получения отклонений от тренда. Однако, регрессионная модель, полученная благодаря таким преобразованиям, слабо поддается интерпретации. Иногда используют априорную информацию об имеющихся между параметрами связях в виде ограничений при вычислении оцениваемых параметров регрессии. За исключением простейших случаев, реализация этих подходов достигается существенным усложнением вычислительной процедуры нахождения оценок. Широкое распространение получили методы устранения мультиколлинеарности, основанные на замене исходного множества объясняющих переменных главными компонентами с последующим отбрасыванием тех из них, которые незначительны в уравнении регрессии. Близким к рассмотренному можно считать методы, основанные не на компонентном, а на факторном анализе, причем аналогия прослеживается как по достоинствам, так и по недостаткам.

В тех случаях, когда перечень объясняющих переменных регрессионной модели слишком велик, рекомендуется разделить их на группы высоко коррелированных и в каждой группе построить обобщающие факторы в виде главной компоненты, которые далее используются как новые переменные строящейся модели.

1.4.2 Доверительные интервалы для уравнения регрессии

Для проведения углубленного анализа уравнения регрессии прежде всего необходимо убедиться в том, что вектор ошибок Е распределен по нормальному закону. Для построения доверительных интервалов коэффициентов модели, предсказанных значений уравнения регрессии, среднего значения используются стандартные статистические распределения, требующие нормальности распределений.

1.4.3 Определение доверительного интервала для истинного значение уравнения регрессии

Определение доверительного интервала сводится к отысканию интервала, в котором с вероятностью содержится истинное значение , соответствующее некоторому опыту из матрицы наблюдений .



Другими словами, имеется интервал, в котором с заданной вероятностью находится линия регрессии.

Подставляя в эмпирическое уравнение регрессии получим оценки для каждого наблюдения вида:



Различие между и объясняется действием различных ошибок.



Отметим, что имеет случайный характер, оценки и распределены нормально с параметрами



,



.



Можно утверждать, что . Другими словами y является состоятельной оценкой истинного значения , соответствующего опыту , т.е. при неограниченном числе опытов эмпирическая линия регрессии совпадает с действительной зависимостью



Составляя дробь Стьюдента, получаем:

.



Задавшись уровнем значимости и найдя табличное значение можно построить достоверный интервал для в виде



.



1.4.4 Свойства доверительных интервалов

а) Доверительный интервал симметричен относительно выборочной оценки ;



б) Ширина доверительного интервала зависит от и ;



в) Ширина доверительного интервала минимальна, если , (ортогональны);



г) Ширина доверительного интервала равна бесконечности, если:

вектор-столбцы и в матрице наблюдений коллинеарные, т.е.если:



д) В общем случае в регрессионных уравнениях доверительный интервал для отдельно взятого регрессионного коэффициента определяется выражением



1.5 Адекватность модели

Существует соотношение, которое можно использовать для оценки адекватности модели, сравнивая и . Расчетное определяется по формуле



(3.4)



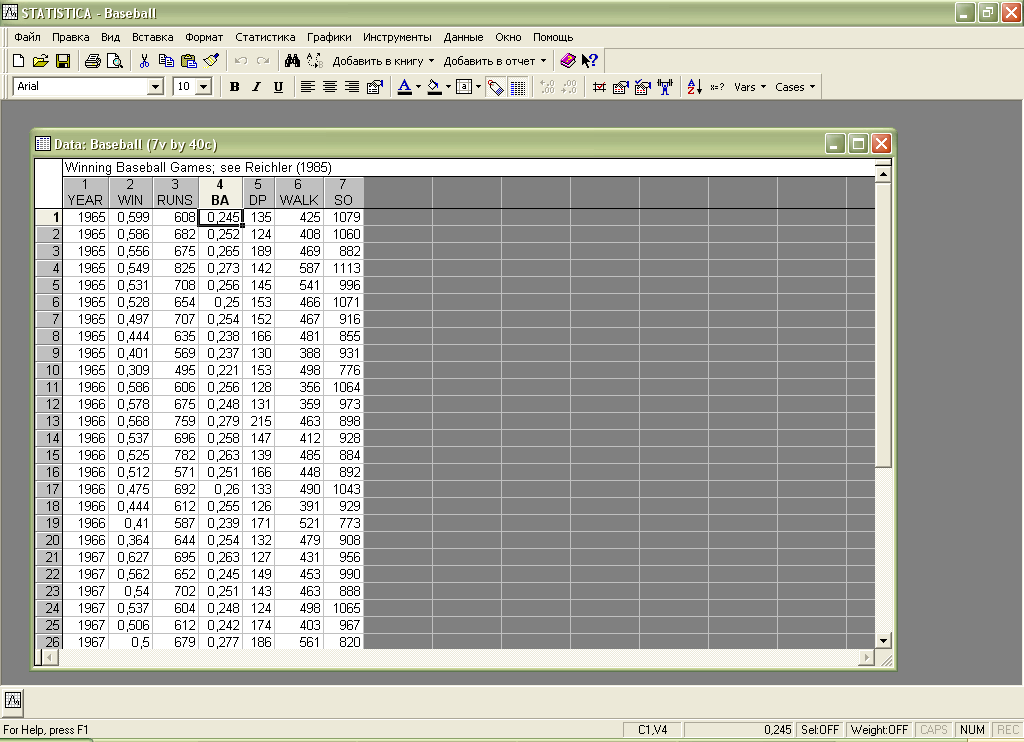
Табличное значение берется с таблиц с определенным числом степенем свобода и для притятого уровня значимості .Если расчетное значение більше , то это значит, что дисперсия MSR статистически меньше дисперсии MSD относительно ,в этом случае полученное уравнение регрессии можно считать дееспособным.



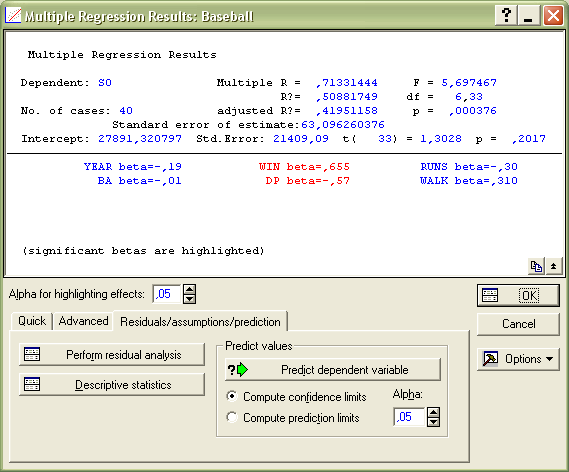
2. ПРАКТИЧЕСКАЯ ЧАСТЬ

Поставлена следующая задача: построить зависимость количества выигранных голов от характеристик сыгранных игр на основе модели множественной регрессии.

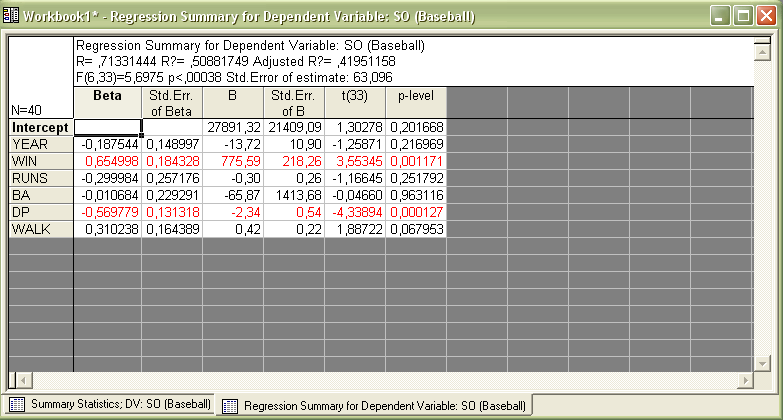
На основе имеющейся выборки сделаем следующие оценки:



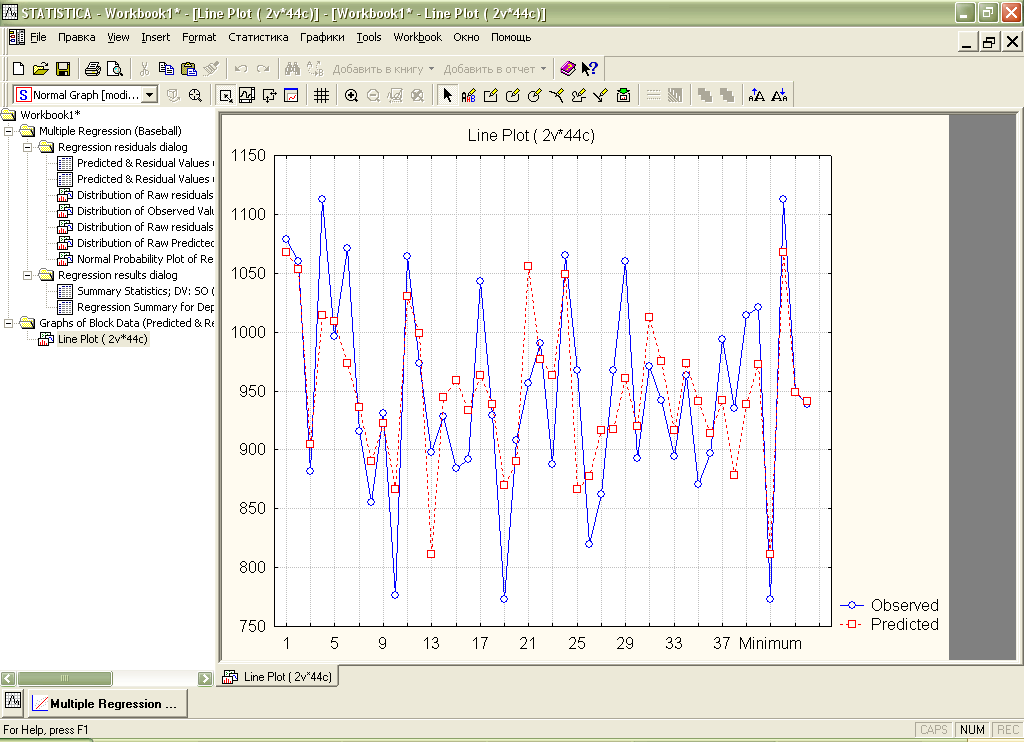
1. параметры модели βi (для данной модели существенными являются переменные WIN и DP):



1. оценки: множественный коэффициент корреляции R, R2 ,F, p, и Std Error of estimate:

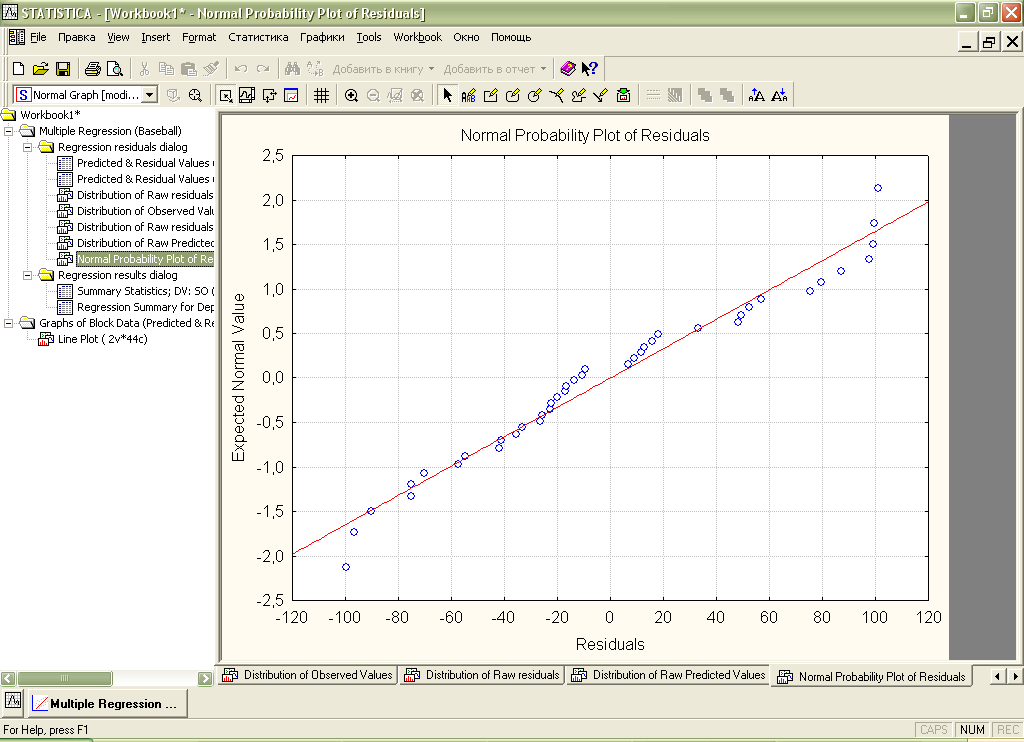


1. график для вычисленных значений и исходных:



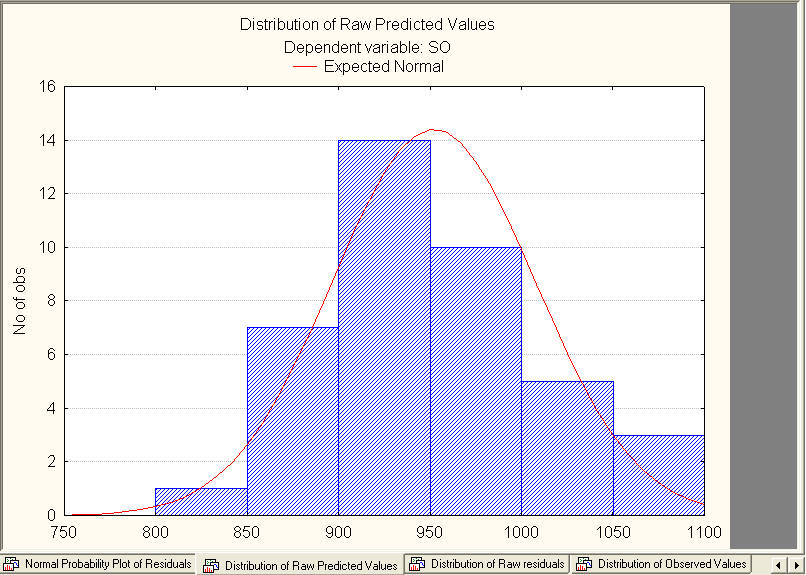
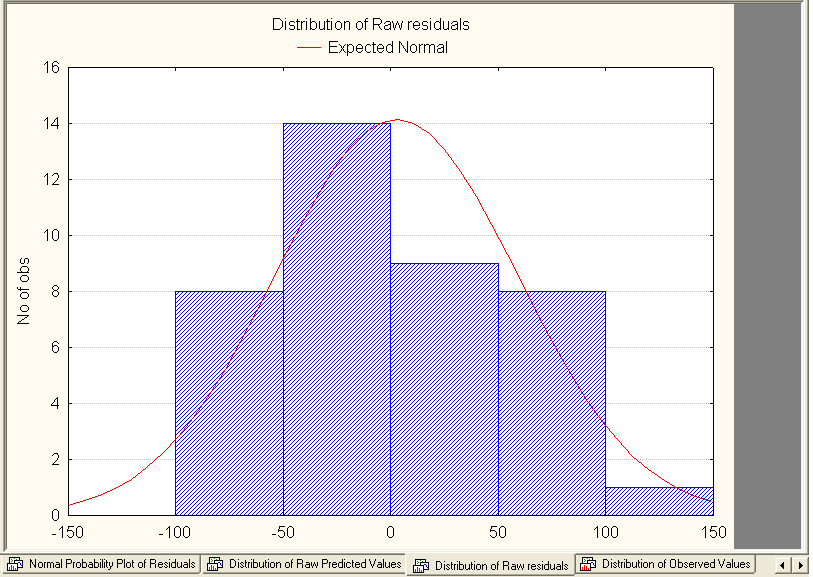
К такому ряду можно применить модель линейной регрессии, так как он стационарный;

1. построение регрессии:



По графику видно, что в целом модель адекватна: практически все значения легли на линию регрессии;

1. гистограммы исходных и вычисленных значений имеют нормальное распределение:



ВЫВОД

Как показано выше, множественная регрессии применима в случае стационарности ряда и позволяет производить мониторинг результатов, основываясь на предикторах.

В общественных и естественных науках процедуры множественной регрессии чрезвычайно широко используются в исследованиях. В общем, множественная регрессия позволяет исследователю задать вопрос (и, вероятно, получить ответ) о том, "что является лучшим предиктором для...". Например, исследователь в области образования мог бы пожелать узнать, какие факторы являются лучшими предикторами успешной учебы в средней школе. А психолога мог быть заинтересовать вопрос, какие индивидуальные качества позволяют лучше предсказать степень социальной адаптации индивида. Социологи, вероятно, хотели бы найти те социальные индикаторы, которые лучше других предсказывают результат адаптации новой иммигрантской группы и степень ее слияния с обществом. Термин "множественная" указывает на наличие нескольких предикторов или регрессоров, которые используются в модели, следовательно такая модель увеличивает спектр анализа регрессоров, что позволит построить более точный прогноз.

ПЕРЕЧЕНЬ ССЫЛОК

1. Демиденко Е.З. Линейная и нелинейная регрессии. – М.: Финансы и статистика, 2010. – 302 с
2. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. – М.: Статистика, 2009. - 437 с.
3. Афифи А., Эйзен С. Статистический анализ. Подход с использованием ЭВМ. Пер. с англ. – М.: Мир, 1982. – 488 с.
4. Тюрин Ю.Н.., Макаров А.А. Статистический анализ данных на компьютере.- М.:Инфра, 1997.-528с.
5. www.statsoft.ru
6. Ясницкий Л.Н. Введение в искусственный интеллект. М. Academia, 2005г.,176 стр.: ил.