

**МЕТОДИЧЕСКИЕ УКАЗАНИЯ**

к выполнению курсовой работы

**разработка статистического программного обеспечения с использованием аналитических моделей**

по дисциплине

«**Интеллектуальный анализ данных**»

для студентов направления подготовки

**09.03.02** *Информационные системы и технологии*

дневной и заочной форм обучения

**Уровень подготовки** *Бакалавр*

#### Севастополь

2015

УДК 519.6

**Разработка статистического программного обеспечения с использованием аналитических моделей.** Методические указания и индивидуальные задания к курсовой работе по дисциплине «Интеллектуальный анализ данных» / Сост. Первухина Е.Л., Токарев А.И. – Севастополь: Изд-во СевГУ, 2015. – 18 с.

Целью курсовой работы является закрепление, углубление и обобщение знаний по специальным разделам математики, необходимых при обработке экспериментальных данных, статистическом моделировании сложных систем различной физической природы.

СОДЕРЖАНИЕ

|  |  |
| --- | --- |
| 1 ЦЕЛЬ И ЗАДАЧИ КУРСОВОЙ РАБОТЫ |  |
| 2 КРАТКИЕ ТЕОРЕТИЧЕСКИЕ СВЕДЕНИЯ |  |

2.1 Корреляционный анализ

2.2 Регрессионный анализ

2.3 F-Фишера.

2.4 Линейный тренд

2.5 Экспоненциальное сглаживание

2.6 Логарифмическое сглаживание

2.7 Степенная функция сглаживания

2.8 ARIMA

2.9 Задания

|  |  |
| --- | --- |
| 3 СОДЕРЖАНИЕ И ОФОРМЛЕНИЕ ПОЯСНИТЕЛЬНОЙ ЗАПИСКИ |  |
| 3.1 Структура пояснительной записки |  |
| 3.2 Содержание разделов пояснительной записки |  |
| БИБЛИОГРАФИЧЕСКИЙ СПИСОК |  |

**1 Цель и задачи курсовой работы**

Целью курсовой работы является закрепление, углубление и обобщение знаний по специальным разделам математики, необходимых при обработке экспериментальных данных, статистическом моделировании сложных систем различной физической природы.

Курсовая работа связана с решением инженерных задач с использованием информационных систем последнего поколения и с приобретением навыков научного исследования, в том числе при написании дипломной работы.

Курсовая работа выполняется по индивидуальным заданиям и включает следующие этапы:

- углубленное изучение методов статистического моделирования сложных систем различной физической природы;

- построение алгоритмов и решение задач;

- разработка аналитического программного обеспечения, используя изученные методы и модели.

В качестве источника экспериментальной информации об экономических системах используется материал базы данных Федеральной службы государственной статистики РФ, Российский статистический ежегодник, реальная экспериментальная информация о параметрах технических объектов.

## 2. КРАТКИЕ ТЕОРЕТИЧЕСКИЕ СВЕДЕНИЯ

## *2.1 Корреляционный анализ*

Показатели корреляции основаны на оценке сопряженной вариации изучаемых переменных. Парный коэффициент корреляции *(r)* – это нормированный коэффициент ковариации. Ковариация, являясь мерой взаимосвязи двух переменных, рассчитывается как средняя величина произведения отклонений индивидуальных значений анализируемых данных от их средних значений:

 (1)

Недостаток этого показателя (числовое значение ковариации зависит от размерности переменных *x* и *y*) преодолевается в парном коэффициенте корреляции путем нормирования абсолютных отклонений.

 и ,

где *σy* – среднее квадратическое отклонение переменной *y*;

*σx* – среднее квадратическое отклонение переменной *x*.

Парный коэффициент корреляции:

, (2)

где *n* – число единиц в статистической совокупности.

Коэффициент корреляции изменяется в пределах

0 *≤ |r| ≤* 1 . (3)

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

## *Множественный корреляционный анализ*

Изучение множественной корреляционной зависимости предполагает оценку влияния на результативный признак двух и более факторов. При этом рассчитываются множественные (совокупные) и частные коэффициенты корреляции, которые можно определить на основе парных коэффициентов корреляции .

Так, при двухфакторной модели связи рекуррентная формула множественного коэффициента корреляции выглядит следующим образом:

. (4)

Значения *R* изменяются в пределах от 0 до 1. Величина совокупного коэффициента корреляции всегда больше любого из парных коэффициентов и включение в анализ новых факторов не может привести к уменьшению значения *R*.

Квадрат множественного коэффициента корреляции *R2* является множественным коэффициентом детерминации и характеризует долю дисперсии результативного признака, объясненную вариацией всех факторов, включенных в анализ, в общей дисперсии результата.

Частные коэффициенты также могут быть рассчитаны по рекуррентной формуле. Если элиминируется влияние одного фактора, то частный коэффициент корреляции называется коэффициентом первого порядка и в условиях двухфакторной модели рассчитывается следующим образом:

, (5)

где – частный коэффициент корреляции первого порядка; точка между *х*1 и *х*2 означает, что элиминируется влияние *х*2.

Если элиминируется влияние фактора *х*1, то частный коэффициент корреляции рассчитывается следующим образом:

. (6)

## *2.2 Парный регрессионный анализ*

Парной регрессией называется уравнение связи двух переменных *x* и *y* вида

, (7)

где *у* – зависимая переменная (результативный признак); *х* – независимая, объясняющая переменная (признак-фактор).

Парная регрессия применяется, если имеется доминирующий фактор, который и используется в качестве объясняющей переменной.

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. Представленное ниже раздельное их описание необходимо для понимания сути каждой процедуры.

Первый этап регрессионного анализа – поиск линии регрессии, которая бы лучшим образом аппроксимировала поле корреляции. При этом необходимо учитывать природу изучаемых показателей, специфику их взаимосвязи, свойства математических функций. Однако в настоящее время процедура выбора лучшего уравнения связи формализована. 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) – прямая, линейная функция;

2)  – парабола;

3)  – гипербола;

4)  – степенная функция;

5)  – 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. Рассмотрим экспоненциальную и степенную функции после линеаризации:

; .

Простейшим видом уравнения регрессии является парная линейная регрессия

,

где – расчетное, теоретическое значение признака-результата; ** – 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.

Чтобы воспользоваться уравнением регрессии, необходимо рассчитать значения его параметров. Чаще всего расчет осуществляется по методу наименьших квадратов (МНК). Суть метода в том, что удается получить такие значения параметров, при которых минимизируется сумма квадратов отклонений эмпирических значений признака-результата от его теоретических значений, т.е.

. (8)

Третий этап регрессионного анализа предполагает оценку качества полученного уравнения связи. Поскольку уравнение регрессии строится, как правило, на основе выборочных данных, то следует оценить статистическую значимость параметров уравнения и уравнения в целом.

Оценка статистической значимости параметров модели означает проверку нулевых гипотез о равенстве параметров генеральной совокупности нулю, т.е. в условиях парной регрессии:

*Н0: =0, Н0: =0.*

Проверка производится с использованием *t*-статистики, которая в этом случае представляет собой отношение значения параметра к его стандартной (среднеквадратической) ошибке *S*:

 и , (9)

поскольку*=*0 и *=*0, то

;, (10)

где - стандартная ошибка параметра: =;

- стандартная ошибка параметра : =.

Фактические значения *t*-критерия сравниваются с табличными (с учетом уровня значимости *α* и числа степеней свободы (*d.f.=n-k-*1)). Параметры признаются статистически значимыми, т.е. сформированными под воздействием неслучайных факторов, если *t*факт > *t*табл.

Результаты оценки регрессионного уравнения могут быть разными. Возможен вариант, когда уравнение в целом статистически значимо, а некоторые параметры уравнения незначимы. Это означает, что описанная зависимость результата от аргументов может служить основой для принятия некоторых управленческих решений, но полученное уравнение регрессии нельзя использовать для прогнозирования. Уравнение связи признается моделью и может быть использовано в целях прогнозирования, если статистически значимы и параметры, и уравнение в целом.

## *Множественный регрессионный анализ*

Модель множественной регрессии в общем виде записывается следующим образом:

 (11)

Специфической проблемой, решаемой при построении множественной регрессии, является отбор факторов, включаемых в уравнение регрессии. Напомним, что для получения надежных оценок параметров необходимо, чтобы число факторов, включаемых в модель, было по меньшей мере в 5 – 6 раз меньше объема изучаемой совокупности. Из-за ограниченности объема совокупности не стоит «засорять» модель факторами, связь которых с зависимой переменной слабая (*r* < 0,3).

Анализ первой строки матрицы, содержащей показатели тесноты связи между признаком-результатом и каждым из признаков-факторов, позволяет исключить из анализа факторы, практически не влияющие на поведение зависимой переменной (*r* < 0,3).

В остальных клетках выделенного сегмента матрицы содержатся коэффициенты, оценивающие зависимость между факторами. Анализ этих характеристик позволяет выявить наличие мультиколлинеарности. Из двух коллинеарных факторов (*r* ≥ 0,7) один следует исключить из анализа. Предпочтение отдается признаку, связь которого с результатом более тесная.

Расчет параметров уравнения множественной регрессии осуществляется на основе МНК.

Параметры при факторах в уравнении множественной регрессии называются условно-чистыми коэффициентами регрессии. Их можно было бы назвать «чистыми», если бы удалось включить в модель все факторы, определяющие значение признака-результата, что на практике не может быть реализовано.

Условно-чистые коэффициенты регрессии оценивают силу влияния каждой независимой переменной при условии элиминирования других факторов, включенных в модель. Интерпретация значений коэффициентов аналогична интерпретации коэффициентов в уравнении парной регрессии.

Независимые переменные могут иметь разные единицы измерения, поэтому получаемые коэффициенты регрессии не сопоставимы и не позволяют ранжировать аргументы по силе их влияния на зависимую переменную.

***2.3 F-Фишера***

Значимость уравнения в целом оценивается на основе *F*-критерия Фишера.

*F*-критерий – это отношение объясненной вариации (факторной дисперсии) результативного признака, рассчитанной на одну степень свободы, , к необъясненной вариации (остаточной дисперсии) признака-результата, рассчитанной на одну степень свободы,, таким образом

:  , (12)

где *k* – число степеней свободы факторной дисперсии, равное числу независимых переменных (признаков-факторов) в уравнении регрессии;

*n-k-*1 - число степеней свободы остаточной дисперсии.

Если обе части соотношения разделить на общую дисперсию зависимой переменной (результата), то *F*-критерий может быть представлен следующим образом:

. (13)

Расчетное значение критерия сопоставляется с табличным (с учетом числа степеней свободы: *d.f. = k* и *d.f.=n-k-*1) (приложение 3). Если , то делается вывод о статистической значимости уравнения в целом. Поскольку *F*-критерий основан на соотношении факторной и остаточной дисперсий результативного признака, то вполне логично его использование для оценки качества модели. Если объясненная дисперсия существенно больше необъясненной, это означает, что в уравнение связи включены именно те факторы, которые играют определяющую роль в изменении значения признака-результата. Статистическая значимость уравнения одновременно означает статистическую значимость коэффициента детерминации.

***2.4 Линейный тренд***

Как и любая другая регрессия, тренд может быть как линейным (степень влияющего фактора t равна 1) так и нелинейным (степень больше или меньше единицы). Так как линейная регрессия является самой простейшей, хотя далеко не всегда самой точной, то рассмотрим более детально именно этот тип тренда.

Общий вид уравнения линейного тренда:

***Y(t) = a0 + a1\*t + Ɛ***

Где a0 – это нулевой коэффициент регрессии, то есть, то каким будет Y в случае, если влияющий фактор будет равен нулю, a1 – коэффициент регрессии, который выражает степень зависимости исследуемого показателя Y от влияющего фактора t, Ɛ – случайная компонента или стандартная ошибка, по сути являет собой разницу между реально существующими значениями Y и расчетными. t – это единственный влияющий фактор – время.

Чем более выраженная тенденция роста показателя или его падения, тем будет больше коэффициент a1. Соответственно, предполагается, что константа a0 совместно со случайной компонентой Ɛ отражают остальные регрессионные влияния, помимо времени, то есть всех прочих возможных влияющих факторов.

Рассчитать коэффициенты модели можно стандартным Методом наименьших квадратов (МНК).

***2.5 Экспоненциальное сглаживание***

Данный метод – один из простейших и наиболее распространённых методов выравнивания временного ряда. Экспоненциальное сглаживание можно представить как фильтр, на вход которого последовательно поступают члены исходного ряда, а на выходе формируются текущие значения экспоненциальной средней. Пусть [image154.png](https://university.prognoz.ru/biu/ru/Image:image154.png) – исходный временной ряд, тогда простое экспоненциальное сглаживание определяется рекуррентной формулой:

[image155.png](https://university.prognoz.ru/biu/ru/Image:image155.png), где

[image140.png](https://university.prognoz.ru/biu/ru/Image:image140.png) – сглаженный ряд (экспоненциальное среднее);

[image139.png](https://university.prognoz.ru/biu/ru/Image:image139.png) – значение временного ряда в момент t;

[image162.png](https://university.prognoz.ru/biu/ru/Image:image162.png) – коэффициент сглаживания.

Когда эта формула применяется рекурсивно, то каждое новое сглаженное значение вычисляется как взвешенное среднее текущего наблюдения из сглаженного ряда. Очевидно, результат сглаживания зависит от параметра [image162.png](https://university.prognoz.ru/biu/ru/Image:image162.png). Если [image156.png](https://university.prognoz.ru/biu/ru/Image:image156.png), то предыдущие наблюдения полностью игнорируются. Если [image157.png](https://university.prognoz.ru/biu/ru/Image:image157.png), то игнорируются текущие наблюдения. Значения [image158.png](https://university.prognoz.ru/biu/ru/Image:image158.png) дают промежуточные результаты.

Если последовательно использовать данное рекуррентное соотношение, то экспоненциальную среднюю [image140.png](https://university.prognoz.ru/biu/ru/Image:image140.png) можно выразить через значения временного ряда [image021.png](https://university.prognoz.ru/biu/ru/Image:image021.png):

[image159.png](https://university.prognoz.ru/biu/ru/Image:image159.png) .

Таким образом, величина [image160.png](https://university.prognoz.ru/biu/ru/Image:image160.png) оказывается взвешенной суммой всех членов ряда. Причём веса уменьшаются экспоненциально в зависимости от давности наблюдения относительно момента времени [image161.png](https://university.prognoz.ru/biu/ru/Image:image161.png). Это и объясняет, почему величина [image160.png](https://university.prognoz.ru/biu/ru/Image:image160.png) названа экспоненциальной средней.

***2.6 Логарифмический тренд***

Логарифмический тренд – это функция y(x)=a\*ln(x)+b,

где

Значение x – это номера периода во временном ряду (например, номер месяца, квартала, дня;[См. статью о временных рядах](http://www.4analytics.ru/chto-vajno-znat-o/o-vremennix-ryadax.html).)

y – это последовательность значений , которые мы анализируем и прогнозируем (например, объём продаж по месяцам.)

b – точка пересечения с осью y на графике;

a – это значение, на которое увеличивается следующее значение временного ряда;

Причем, если a>0, то динамика роста положительная,

Если а<0, то динамика тренда отрицательная.

При построении логарифмического тренда используют как положительные, так и отрицательные данные временных рядов.

***2.7 Степенной тренд***

**Степенные тренды** используются, когда данные состоят из результатов измерений, значения которых плавно увеличиваются с нарастающей скоростью. При этом данные не могут содержать нулевых и отрицательных значений.

Степенной тренд определяется следующей формулой:

[image120.png](https://university.prognoz.ru/biu/ru/Image:image120.png)

***2.8 ARIMA***

В большинстве временных рядов члены ряда зависят друг от друга. На свойстве влияния предыдущего состояния процесса на будущее базируются модели авторегрессии. Математически это свойство можно выразить уравнением

,

где  – значение  в момент времени ;  – коэффициенты уравнения (*i*= 1,2,…,р); *р* – порядок авторегрессии;  – случайная величина.

В модели скользящего среднего в отличие от предыдущего способа предполагают, что каждый элемент ряда подвержен суммарному воздействию случайных предыдущих ошибок :

,

где  – значение  в момент времени ;  – коэффициенты уравнения (*i*=1,2,…,q); *q* – порядок модели скользящего среднего; – случайная величина.

***2.9 Задания***

**Выбор варианта выполнения задания остается за студентом**

**Вариант выполнения задания №1:**

1. Написать оконное аналитическое приложение. Задача - анализ временных рядов (данные по вариантам), составление трендов и прогнозов с использованием различных моделей.

2. Приложение должно произвести загрузку данных из файла (или из Интернета).

3. В наличии должна быть функция ввода данных непосредственно в само приложение.

4. Необходимо, чтоб приложение могло сохранять результаты исследований, а также загружать сохраненные файлы.

5. При проведении корреляционно-регрессионного анализа в приложении должен автоматически рассчитываться F-критерий Фишера и сравниваться с табличным значением.

6. При построении линий тренда, автоматически должен рассчитываться коэффициент достоверности аппроксимации (.

7. По его результатам должна строиться наилучшая модель из предложенных по вариантам.

8. При построении прогнозов автоматически должны строиться доверительные интервалы.

9. Дальность прогноза задается пользователем.

10. Должен выводиться график и формула модели, с помощью которой мы проводили исследование временных рядов.

11. Право выбора модели должно оставаться за пользователем, но также необходима функция автоматического выбора модели (в случаях, где можно использовать коэффициент достоверности аппроксимации).

12. В приложениях к пояснительной записке необходимо включить исходный код в текстовом файле, скриншоты работы программы и записанную программу на диске.

**Вариант выполнения задания №2:**

1. Получить варианты исходных данных, не менее 60 показателей (временные ряды с конкретными показателями, а именно нефть, золото и т.п.)

2. Для выполнения курсовой работы использовать следующие статистические программы: Statistica, Statgraphics и Minitab.

3. Сравнить в этих трех программах проведение корреляционного, регрессионного анализа, линейного, экспоненциального тренда, степенной функции сглаживания, логарифмического сглаживания. Так же сравнить построение моделей ARIMA. (Это задание для всех вариантов одно)

4. **В теоретической части** должны быть подробно описаны формулы и способы проведения всех вышеперечисленных видов анализа.

5. Должна быть следующая структура **практической части** курсовой работы: Берутся определенные данные (по выбору), проводится какой-то один вид анализа, к примеру, корреляционный, но при этом он проводится в трех статистических программах. После проведения какого-либо анализа, в пояснительной записке пишется содержательный вывод, прикрепляются скриншоты программ и графики. Так же в написании выводов должны быть указаны особенности работы каждой из программ, их сходства и различия.

**3 СОДЕРЖАНИЕ И ОФОРМЛЕНИЕ ПОЯСНИТЕЛЬНОЙ ЗАПИСКИ**

Заключительным этапом выполнения работы являются оформление пояснительной записки, графического материала и защита студентом курсовой работы в соответствии с утвержденным индивидуальным заданием.

**3.1 Структура пояснительной записки**

Пояснительная записка к курсовому проекту выполняется на стандартных листах формата А4 с одной стороны листа и должна иметь следующую структуру:

|  |  |
| --- | --- |
| 1 | Титульный лист |
| 2 | Индивидуальное задание на выполнение работы |
| 3 | Содержание |
| 4 | Введение |
| 5 | Постановка задачи |
| 6 | Содержательная часть работы по отдельным разделам индивидуального задания |
| 7 |
| 8 |
| 9 |
| 10 |
| 11 | Заключение |
| 12 | Библиографический список |
| 13 | Приложения |

**3.2 Содержание разделов пояснительной записки**

Титульный лист является первым листом пояснительной записки, с него начинается нумерация страниц, но номер 1 на титульном листе не ставится. Образец титульного листа приведен в приложении А.

Индивидуальное задание на выполнение работы представляет специальный бланк, который необходимо заполнить. Образец задания приведен в приложении Б.

В содержании перечисляют наименования разделов пояснительной записки и номера соответствующих страниц.

Во введении указывают цели и задачи курсового проектирования.

В разделе «Постановка задачи» приводят постановку задачи на курсовое проектирование.

В содержательной части работы по отдельным разделам, соответствующим индивидуальному заданию, описывают алгоритмы решения задач (отмечая их достоинства и недостатки) с основными и промежуточными преобразованиями и подстановкой числовых значений, а также приводят результаты проверки правильности решения.

В разделе “Заключение” должны быть сделаны основные выводы по результатам курсового проектирования.

В разделе "Библиографический список" указывают перечень научно-технических публикаций, нормативных технических документов и других научно-технических материалов, на которые есть ссылки в основном тексте. Если на источник нет ссылки из основного текста, его в список не помещают.Список должен быть приведен в конце текста пояснительной записки, начиная с новой страницы. Библиографические описания в перечне ссылок приводят в порядке, в котором они впервые упоминаются в тексте. Порядковые номера описаний в перечне являются ссылками в тексте (номерные ссылки). Библиографические описания ссылок в перечне приводят в соответствии с действующими стандартами по библиотечному и издательскому делу.

В приложение к пояснительной записке могут быть включены таблицы, обоснования, методики, расчеты, блок-схемы алгоритмов программ, графические материалы и другие документы, использованные при выполнении курсовой работы.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1.Box G.E. Time Series Analysis, Forecasting and Control / G.E. Box., G.M. Jenkins, G.C. Reinsel. — New Jersey: Prentice Hall, 1994. — 598 p.

2. W. Enders. Applied Econometric Time Series / W. Enders. — New York: John Wiley & Sons, 1995 — 433 p.

3. Суслов В.И. Эконометрия / В.И. Суслов, Н.М. Ибрагимов, Л.П. Талышева, А.А. Цыплаков. — Новосибирск: Издательство СО РАН, 2005. — 744с.

4. Айвазян С. А. Прикладная статистика. Основы эконометрики: Учебник для вузов/ С. А. Айвазян, В. С. Мхитарян. М. : ЮНИТИ-ДАНА, 2001. –656 с.

5. Куприенко Н. В. Статистические методы изучения связей. Корреляционно-регрессионный анализ/ Н. В. Куприенко, О. А. Пономарева, Д. В. Тихонов. СПб. : Изд-во политехн. ун-та, 2008. – 118 с.

6. Шанченко, Н. И. Лекции по эконометрике : учебное пособие для студентов высших учебных заведений, обучающихся по специальности «Прикладная информатика (в экономике)» / Н. И. Шанченко. – Ульяновск : УлГТУ, 2008. – 139 с.