Министерство образования и науки Российской Федерации
ФГАОУ ВО «Севастопольский государственный университет»

Институт информационных технологий
и управления в технических системах

Лабораторная работа №2
«Корреляционный и регрессионный анализ данных»

по дисциплине «Интеллектуальный анализ данных»
для студентов всех форм обучения направления подготовки
09.03.02 «Информационные системы и технологии»

Севастополь
2017

Корреляционный и регрессионный анализ данных. Методические указания к лабораторным занятиям по дисциплине «Интеллектуальный анализ данных» / Сост.: И.В. Дымченко, И.П. Шумейко, О.А. Сырых – Севастополь: Изд-во СевГУ, 2017 – 13 с.

Методические указания предназначены для проведения лабораторных работ по дисциплине «Интеллектуальный анализ данных». Целью методических указаний является помощь студентам в изучении возможностей системы RStudio. Излагаются практические сведения необходимые для выполнения лабораторной работы, требования к содержанию отчета.

Методические указания рассмотрены и утверждены на методическом семинаре и заседании кафедры «Информационные системы» (протокол № 1 от 29 августа 2016 г.)

Лабораторная работа №2.4

Корреляционный и регрессионный анализ данных. Множественная линейная регрессия.

Цель:

- исследовать возможности языка R для построения множественной линейной регрессий.

Время: 2 часа

Лабораторное оборудование: персональные компьютеры, выход в сеть Internet, RStudio.

Краткие теоретические сведения

Когда в регрессионной модели есть одна зависимая и одна независимая переменная, такой подход называется простой линейной регрессией. Когда есть одна зависимая переменная, но в модель входят ее степени (например, X, X^2, X^3), это называется полиномиальной регрессией. Если есть больше одной независимой переменной, это называется множественной регрессией.

Множественная линейная регрессия

Если существует больше одной независимой переменной, простая линейная регрессия превращается во множественную линейную регрессию, а ход вычислений становится более сложным.

Множественная линейная регрессия позволяет изучить совместное воздействие нескольких независимых переменных на переменную отклика. Практическое применение двоякое: для предсказания переменной отклика и для определения интенсивности, с которой каждая независимая переменная линейно связана с зависимой. В общем случае уравнение множественной линейной регрессии имеет вид:

$$y = b_0 + b_1x_1 + b_2x_2 ... + b_kx_k .$$

Если существует больше одной независимой переменной, то регрессионные коэффициенты показывают, на сколько увеличится значение зависимой переменной при изменении данной независимой переменной на единицу при условии, что все остальные независимые переменные останутся неизменными.

Важным дополнительным условием является некоррелированность объясняющих переменных между собой (отсутствие мультиколлинеарности). Считается, что явление мультиколлинеарности наблюдается тогда, когда коэффициент корреляции между объясняющими переменными превышает по модулю 0,7.

Пошаговая множественная линейная регрессия

Существуют две методики построения множественной регрессии – пошаговая вперед и пошаговая назад.

Пошаговая вперед заключается в то, что первоначально строится модель с одной экзогенной переменной. Затем добавляется следующая и строится новая модель. Модели сравниваются и, в зависимости от того ухудшилась или улучшилась модель, введенная переменная либо остается в модели, либо заменяется на другую. Таким образом, перебираются различные комбинации экзогенных переменных, в результате получается наилучшая модель.

Пошаговая назад начинается с того, что рассчитывается множественная регрессия на всем множестве факторов. Затем построенная модель исследуется с точки зрения статистической значимости модели в целом, статистической значимости коэффициентов регрессии, оценивается коэффициент детерминации. Затем из модели удаляется один из влияющих факторов.

Его выбор можно осуществить следующим образом:

1. Строится матрица парных коэффициентов корреляции между переменными.

2. Выбираются две экзогенные переменные, между которыми наибольший коэффициент парной корреляции.

3. Из этих двух переменных выбирается та, которая оказывает меньшее влияние на эндогенную переменную, и исключается из модели.

Затем строится новая модель, исследуется ее качество. Также проводится тест на лучшую из двух моделей: с меньшим или большим числом переменных. В конце получается наилучшая модель.

Результат применения метода пошаговой регрессии зависит от критериев включения или удаления переменных. При помощи функции `stepAIC()` из пакета MASS можно провести построение пошаговой регрессии с использованием точного критерия AIC (Akaike Information Criterion – информационный критерий Акаике).

При расчете этого критерия учитывается статистическое соответствие модели данным и число необходимых для достижения этого соответствия параметров. Предпочтение нужно отдавать моделям с **меньшими** значениями AIC, указывающими на хорошее соответствие данным при использовании меньшего числа параметров**.**

Мультиколлинеарность

Наибольшие затруднения в использовании аппарата множественной регрессии возникают при наличии мультиколлинеарности факторных переменных, когда более чем два фактора связаны между собой линейной зависимостью.

Мультиколлинеарностью для линейной множественной регрессии называется наличие линейной зависимости между факторными переменными, включёнными в модель.

Мультиколлинеарность – нарушение одного из основных условий, лежащих в основе построения линейной модели множественной регрессии.

Мультиколлинеарность можно выявить на начальном этапе моделирования (до построения регрессии). О ней могут свидетельствовать:

1. Большие (по абсолютной величине) парные коэффициенты корреляции между независимыми переменными.

2. Высокие (>10) значения коэффициента *VIF*.

Коэффициент *VIF* (variance inflation factor) характеризует силу мультиколлинеарности. Вычисление коэффициента выполняется с помощью функции `vif()`

Симптомами присутствия мультиколлинеарности в уже построенной модели являются:

1. Небольшое изменении исходных данных, приводит к существенному изменению оценок коэффициентов.

2. Каждая переменная в отдельности является незначимой, а уравнение в целом имеет высокий R^2 (коэффициент детерминации) и является значимым.

3. Оценки коэффициентов имеют неправильные с точки зрения экономической теории знаки или неоправданно большие значения

Задание и порядок выполнения лабораторной работы №2.4

Построение множественной линейной регрессии

1. Запустить пакет R commander (Rcmdr).
2. Загрузить данные для анализа из файла Данные.xlsx
3. Исследуйте парные взаимосвязи между всеми переменными.
4. Постройте диаграммы рассеяния
5. Проведите подгонку множественной регрессионной модель при помощи функции lm(). Для этого выберете следующие пункты меню в оболочке R commander:

Статистики

Подгонка моделей

Линейная регрессия

В появившемся диалоговом окне укажите: зависимая переменная – индекс реального ВВП, независимые – остальные индексы. По таблице коэффициентов запишите полученное уравнение регрессии, проанализируйте полученные коэффициенты регрессии.

Обозначения

• Первый столбец (без заголовка) – свободный член (`intercept`) и предикторы в модели;

• `Estimate` – коэффициент регрессии;

• `Std. Error` – стандартная ошибка коэффициента регрессии;

• `t value` – t-статистика, с помощью которой проверяют статистическую гипотезу отличия коэффициента регрессии от 0;

• `Pr(>|t|)` – уровень значимости предиктора, статистически значимые предикторы помечены звёздочками;

• `Residual standard error` – остаточная стандартная ошибка – показатель разброса возможных значений случайной ошибки;

• `Multiple R-squared` – коэффициент детерминации, или квадрат коэффициента множественной корреляции модели;

• `Adjusted R-squared` –скорректированный коэффициент детерминации;

• `F-statistic` – F-статистика, оценивающая значимость полученной модели множественной линейной регрессии в целом, и её уровень значимости.

6. Проанализируйте графики остатков (наблюдаемое минус предсказанное регрессионной моделью значение).

Модели

Графики

График Компонента + остаток

7. Загрузите свои экспериментальные данные.

8. Постройте множественную линейную регрессию. По таблице коэффициентов запишите полученное уравнение регрессии. Проанализируйте график остатков.

Построение пошаговой множественной регрессии

1. Запустить пакет R commander (Rcmdr).
2. Загрузить данные для анализа из файла Данные.xlsx
3. Выполнить пошаговое построение регрессии :

Модели

Ступенчаты выбор модели

В открывшемся окне выбрать коэффициент AIC, направление вперед

4. Сделать выводы по полученным результатам

5. Аналогично сделать пошаговое построение по направлению назад, сравнить с предыдущим построением, сделать выводы.

6. Проверить коэффициент VIF и сделать выводы

Модели

Числовая диагностика

Факторы влияющие на дисперсию

7. Загрузить свои экспериментальные данные

8. Выполнить пошаговое построение множественной регрессии по направлению назад (вперед) сравнить полученные результаты, провести проверку коэффициент VIF, сделать выводы.

Содержание отчета

Отчет по выполняемой лабораторной работе выполняется каждым студентом индивидуально на листах формата А4 в рукописном или машинном варианте исполнения и должен содержать:

- название работы;
- цель и задачи исследований;
- набор экспериментальных данных;
- выводы по построенной множественной регрессии;
- программный код с комментариями;
- выводы по работе.

Контрольные вопросы

1. Множественная регрессия
2. Пошаговая множественная регрессия
3. Мультиколлинеарность.

Библиография

1. Алексей Шипунов и др. Наглядная статистика. Используем R! – М.: ДМК Пресс, 2014. – 298 с. [Электронный ресурс]. Режим доступа: http://ashipunov.info/shipunov/school/books/rbook.pdf.
2. Зарядов И.С. Введение в статистический пакет R: типы переменных, структуры данных, чтение и запись информации, графика. М.: Издательство Российского университета дружбы народов, 2010. – 207 с.
3. Роберт И. Кабаков R в действии. Анализ и визуализация данных в программе R / пер. с англ. Полины А. Волковой. – М.: ДМК Пресс, 2014. – 588 с.: ил.
4. Оффициальный сайт RStudio. Режим доступа: https://www.rstudio.com.
5. Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных [Электронный ресурс]. Режим доступа: http://machinelearning.ru.
6. Мастицкий С.Э., Шитиков В.К. (2014) Статистический анализ и визуализация данных с помощью R. – Электронная книга. Режим доступа: http://r-analytics.blogspot.com

Организация защиты и критерии оценивания выполнения лабораторных работ

К защите представляется отчет, включающий в себя результаты выполнения лабораторной работы, выполненный согласно правилам и единый титульный лист, на котором отмечаются результаты выполнения заданий.

К отчетам прилагается электронный носитель, содержащий папки с исполняемыми файлами, файлами отчетов и презентациями (если требуется в задании) созданных в ходе выполнения лабораторных работ.

На проверку теоретической подготовки, проводимой по контрольным вопросам, отводиться 5–6 минут.

Степень усвоения теоретического материала оценивается по следующим критериям:

- ***оценка «отлично» выставляется, если:***
  - последовательно, четко, связно, обоснованно и безошибочно с использованием принятой терминологии изложен учебный материал, выделены главные положения, ответ подтвержден конкретными примерами, фактами;
  - самостоятельно и аргументировано сделан анализ, обобщение, выводы, установлены межпредметные (на основе ранее приобретенных знаний) и внутрипредметные связи, творчески применены полученные знания в незнакомой ситуации;
  - самостоятельно и рационально используются справочные материалы, учебники, дополнительная литература, первоисточники; применяется систему условных обозначений при ведении записей, сопровождающих ответ; используются для доказательства выводы из наблюдений и опытов, ответ подтверждается конкретными примерами;
  - допускает не более одного недочета, который легко исправляется по требованию преподавателя.
- ***оценка «хорошо» ставится, если:***
  - дан полный и правильный ответ на основе изученных теорий; допущены незначительные ошибки и недочеты при воспроизведении изученного материала, определения понятий, неточности при использовании научных терминов или в выводах и обобщениях из наблюдений и опытов; материал излагает в определенной логической последовательности;
  - самостоятельно выделены главные положения в изученном материале; на основании фактов и примеров проведено обобщение, сделаны выводы, установлены внутрипредметные связи.
  - допущены одна негрубая ошибку или не более двух недочетов, которые исправлены самостоятельно при требовании или при небольшой помощи преподавателя; в основном усвоил учебный материал.
- ***оценка «удовлетворительно» ставится, если:***
  - усвоено основное содержание учебного материала, но имеются пробелы в усвоении материала, не препятствующие дальнейшему изучению; материал излагает несистематизированно, фрагментарно, не всегда последовательно;
  - показана недостаточная сформированность отдельных знаний и умений; выводы и обобщения аргументируются слабо, в них допускаются ошибки;
  - допущены ошибки и неточности в использовании научной терминологии, даются недостаточно четкие определения понятий; в качестве доказательства не используются выводы и обобщения из наблюдений, фактов, опытов или допущены ошибки при их изложении;

- обнаруживается недостаточное понимание отдельных положений при воспроизведении текста учебника (записей, первоисточников) или неполные ответы на вопросы преподавателя, с допущением одной – двух грубых ошибок.

- ***оценка «неудовлетворительно» ставится, если:***
  - не усвоено и не раскрыто основное содержание материала; не сделаны выводы и обобщения;
  - не показано знание и понимание значительной или основной части изученного материала в пределах поставленных вопросов или показаны слабо сформированные и неполные знания и неумение применять их к решению конкретных вопросов и задач по образцу;
  - при ответе (на один вопрос) допускается более двух грубых ошибок, которые не могут быть исправлены даже при помощи преподавателя;
  - не даются ответы ни на один их поставленных вопросов.

Оценка выполнения лабораторных работ проводится по следующим критериям

- ***оценка «отлично» ставится, если студент:***
  - творчески планирует выполнение работы;
  - самостоятельно и полностью использует знания программного материала;
  - правильно и аккуратно выполняет задание;
  - умеет пользоваться литературой и различными информационными источниками;
  - выполнил работу без ошибок и недочетов или допустил не более одного недочета
- ***оценка «хорошо» ставится, если студент:***
  - правильно планирует выполнение работы;
  - самостоятельно использует знания программного материала;
  - в основном правильно и аккуратно выполняет задание;
  - умеет пользоваться литературой и различными информационными источниками;
  - выполнил работу полностью, но допустил в ней: не более одной негрубой ошибки и одного недочета или не более двух недочетов.
- ***оценка «удовлетворительно» ставится, если студент:***
  - допускает ошибки при планировании выполнения работы;
  - не может самостоятельно использовать значительную часть знаний программного материала;
  - допускает ошибки и неаккуратно выполняет задание;
  - затрудняется самостоятельно использовать литературу и информационные источники;
  - правильно выполнил не менее половины работы или допустил:
    - не более двух грубых ошибок или не более одной грубой и одной негрубой ошибки и одного недочета;
    - не более двух– трех негрубых ошибок или одной негрубой ошибки и трех недочетов;
    - при отсутствии ошибок, но при наличии четырех–пяти недочетов.
- ***оценка «неудовлетворительно» ставится, если студент:***
  - не может правильно спланировать выполнение работы;
  - не может использовать знания программного материала;
  - допускает грубые ошибки и неаккуратно выполняет задание;
  - не может самостоятельно использовать литературу и информационные источники;
  - допустил число ошибок недочетов, превышающее норму, при которой может быть выставлена оценка «3»;
  - если правильно выполнил менее половины работы;
  - не приступил к выполнению работы;
  - правильно выполнил не более 10% всех заданий.