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 №2.3
Корреляционный и регрессионный анализ данных. Исследование тесноты взаимосвязей данных в среде R

Цель:

- исследовать возможности языка R для определения тесноты взаимосвязей экспериментальных данных;

Время: 2 часа

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

Краткие теоретические сведения

Процедура поиска предполагаемой зависимости между различными числовыми совокупностями обычно включает следующие этапы:

− установление значимости связи между ними;

− возможность представления этой зависимости в форме математиче-ского выражения (уравнения регрессии).

Первый этап в указанном статистическом анализе касается выявления так называемой корреляции, или корреляционной зависимости. Корреляция рассматривается как признак, указывающий на взаимосвязь ряда числовых последовательностей. Иначе говоря, корреляция характеризует силу взаимосвязи в данных. Если это касается взаимосвязи двух числовых массивов x_i и y_i, то такую корреляцию называют парной.

При поиске корреляционной зависимости обычно выявляется вероятная связь одной измеренной величины x (для какого-то ограниченного диапазона ее изменения, например от x_1 до x_n) с другой измеренной величиной y (также изменяющейся в каком-то интервале $y_1 \ldots y_n$). В таком случае мы будем иметь дело с двумя числовыми последовательностями, между которыми и надлежит установить наличие статистической (корреляционной) связи. На этом этапе пока не ставится задача определить, является ли одна из этих случайных величин функцией, а другая – аргументом. Отыскание количественной зависимости между ними в форме конкретного аналитического выражения $y = f(x)$ − это задача уже другого анализа, регрессионного.

Таким образом, корреляционный анализ позволяет сделать вывод о силе взаимосвязи между парами данных x и y, а регрессионный анализ используется для прогнозирования одной переменной (y) на основании другой (x). Иными словами, в этом случае пытаются выявить причинно-следственную связь между анализируемыми совокупностями. Схематическое изображение изложенных соображений представлено на рис.1.

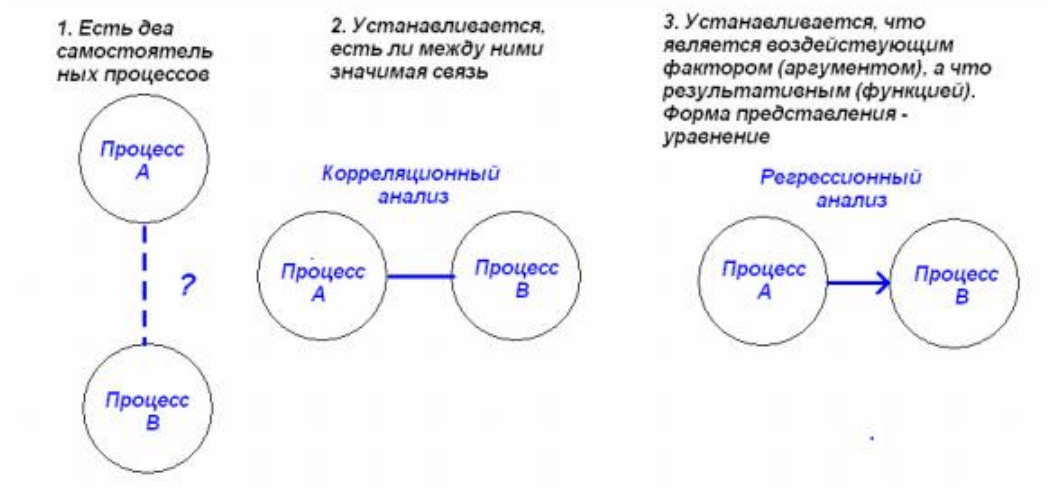

Рисунок.1. Схематическое пояснение сути корреляционного и регрессионного анализов

Принято различать два вида связи между числовыми совокупностями – это может быть **функциональная** зависимость или же **статистическая** (случайная). При наличии функциональной связи каждому значению воздействующего фактора (аргумента) соответствует строго определенная величина другого показателя (функции), т.е. изменение результативного признака всецело обусловлено действием факторного признака.

По своему характеру корреляционные связи – это соотносительные связи.

Такая зависимость графически изображается в виде экспериментальных точек, образующих поле рассеяния, или, как принято говорить, поле корреляции (рис.2).

Рисунок 2. Поле корреляции.

Следовательно, такие двумерные данные можно анализировать с использованием диаграммы рассеяния в координатах «x – y», которая дает визуальное представление о взаимосвязи исследуемых совокупностей. Для количественной оценки существования связи между изучаемыми совокупностями случайных величин используется специальный статистический показатель – **коэффициент корреляции** r.

Если предполагается, что эту связь можно описать линейным уравнением типа $y = a + bx$ (где a и b − константы), то принято говорить о существовании линейной корреляции. Коэффициент r − это безразмерная величина, она может меняться от 0 до ±1. Чем ближе значение коэффициента к единице (неважно, с каким знаком), тем с большей уверенностью можно утверждать, что между двумя рассматриваемыми совокупностями переменных существует линейная связь. Иными словами, значение какой-то одной из этих случайных величин (y) существенным образом зависит от того, какое значение принимает другая (x). Если окажется, что $r = 1$ (или -1), то имеет место классический случай чисто функциональной зависимости (т.е. реализуется идеальная взаимосвязь).

При анализе двумерной диаграммы рассеяния можно обнаружить различные взаимосвязи. Простейшим вариантом является линейная взаимосвязь, которая выражается в том, что точки размещаются случайным образом вдоль прямой линии. Диаграмма свидетельствует об отсутствии взаимосвязи, если точки расположены случайно, и при перемещении слева направо невозможно обнаружить какой-либо уклон (ни вверх, ни вниз). Если точки на ней группируются вдоль кривой линии, то диаграмма рассеяния характеризуется нелинейной взаимосвязью.

Методы определения корреляционной связи

Корреляцию и регрессию принято рассматривать как совокупный процесс статистического исследования, поэтому их использование в статистике часто именуют корреляционно-регрессионным анализом.

Чтобы выявить наличие качественной корреляционной связи между двумя исследуемыми числовыми наборами экспериментальных данных, существуют различные методы, которые принято называть элементарными. Ими могут быть приемы, основанные на следующих операциях:

− параллельном сопоставлении рядов;

− построении корреляционной и групповой таблиц;

− графическом изображении с помощью поля корреляции.

Другой метод, более сложный и статистически надежный, − это количественная оценка связи посредством расчета коэффициента корреляции и его статистической проверки.

Корреляция моментов Пирсона. Если форма распределения анализируемых признаков не очень сильно отличается от нормальной и отсутствуют выбросы, рассчитывают коэффициент **корреляции Пирсона** (часто называемый просто коэффициентом корреляции):

$$r_{xy} = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right] \cdot \left[n\sum y^2 - \left(\sum y\right)^2\right]}}$$

где x - значение факторного признака;

y - значение результативного признака;

n - число пар данных.

Коэффициент корреляции величина относительная; он принимает значение от минус единицы до плюс единицы, т.е. $-1 < r < 1$.

При $r > 0$ связь оценивается, как прямая, при $r < 0$ – обратная.

При $r = 0$ – связь отсутствует, при $|r| = 1$ – связь функциональная

Сила связи оценивается:

при $|r| < 0{,}3$ – как слабая,

при $0{,}3 < |r| < 0{,}7$ – умеренная,

при $|r| > 0{,}7$ – сильная.

Следует помнить, что корреляция между величинами x и y не обя-зательно отражает причинно-следственные связи между ними.

Ранговые коэффициенты корреляции. Непараметрические, или ранговые коэффициенты корреляции Спирмена или Кендалла, рассчитывают в следующих случаях:

– форма распределения отличается от нормальной, например, скошена в ту или иную сторону;

– есть значительные выбросы (которые отражают не ошибки измерений или регистрации данных, а их реальные особенности);

– шкала измерений не количественная, а порядковая;

– небольшой размер выборки.

При вычислении коэффициента корреляции **Спирмена** (r_s) величины x сортируют по возрастанию и ранжируют. Равные величины получают средние значения из рангов, которые получили бы эти значения без ограничения. Аналогично присваивают ранги величинам y. Находят разности рангов x_i и y_i, обозначаемые d_i.

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}.$$

Коэффициент корреляции **Кендалла** (τ) рассчитывают по следующему алгоритму:

1. Значения x ранжируют по возрастанию.

2. Значения y располагают соответственно x и ранжируют.

3. Для каждого ранга y_i определяют число следующих за ним значений рангов, больших y_i. Суммируют и получают n_c – число согласованных пар (от англ. *concordant*), или последовательностей.

4. Для каждого значения y_i определяют число следующих за ним рангов, меньших y_i. Суммируют и получают n_d –число рассогласованных пар (от англ. *discordant*), или инверсий.

5. Рассчитывают τ по формуле:

$$\tau = \frac{2(n_c - n_d)}{n(n-1)}$$

Статистическая значимость коэффициента корреляции. Уровень статистической значимости коэффициента корреляции (Пирсона, Спирмена или Кендалла) зависит от числа наблюдений и может быть оценен с помощью следующей статистики:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

где r – коэффициент корреляции, n – число наблюдений.

При этом проверяются следующие статистические гипотезы:

H_0: Корреляция между переменными не отличается от нуля.

H_1: Корреляция между переменными достоверно отличается от нуля.

При уровне статистической значимости $p<0,05$ отвергается нулевая гипотеза и считается, что связь между изучаемыми переменными действительно существует.

Функции R для проведения корреляционного анализа

Функция `cor` ()

Для расчета коэффициентов корреляции применяют функцию:

`cor (x, use=, method= ).`

Функция возвращает матрицу корреляций между указанными переменными.

Аргументы

`x` – набор данных, между которыми вычисляются корреляции (можно указать имена переменных рабочего файла в кавычках);

`use` – обработка пропусков.

Возможные значения:

`all.obs` – без пропусков (пропущенные данные вызовут ошибку);

`complete.obs` –построчная обработка пропусков;

`pairwise.complete.obs` – попарная обработка пропусков;

`method` – вид корреляции. Возможные значения:

`pearson` – корреляция Пирсона;

`spearman` – корреляция Спирмена;

`kendall` – корреляция Кендалла.

Функция `cor.test` ()

Для оценки уровня значимости коэффициентов корреляции применется функция:

`cor.test (x, y, alternative=, method= ).`

Аргументы

`x, y` – набор данных, должны быть одинаковой длины;

`alternative` –альтернативная гипотеза (двусторонний или односторонний тест). Возможные значения:

`two.sided`;

`greater`;

`less`;

`method` – вид тестируемой корреляции. Возможные значения:

`pearson` – корреляция Пирсона;

`spearman` – корреляция Спирмена;

`kendall` – корреляция Кендалла.

Функции R для графической интерпретации корреляционного анализа

Для визуального исследования зависимости между двумя переменными используют двумерные диаграммы рассеяния, или графики разброса.

Функция `scatterplot` ()

Для создания графика разброса между двумя переменными применяют функцию:

`scatterplot(formula, data, xlab, ylab, legend.title, ellipse, reg.line, smooth).`

Аргументы

`formula` – «формула» для построения графика, применяют в форме y~x или y~x | z, где z – фактор, подразделяющий выборку на подгруппы;

`data` – массив данных, по которому строится график разброса;

`xlab` – название горизонтальной оси;

`ylab` – название вертикальной оси;

`legend.title` – заголовок легенды;

`ellipse` – при значении TRUE вместо точек на графике отображаются корреляционные эллипсы;

reg.line – отображает линию линейной регрессии при значении TRUE и не отображает её при значении FALSE;

smooth – отображает кривую нелинейной регрессии при значении TRUE и не отображает её при значении FALSE.

При необходимости построить матрицу парных графиков по нескольким переменным можно воспользоваться функцией: `scatterplot.matrix ().`

Аргументы данной функции во многом аналогичны таковым функции `scatterplot`(), иначе пишется «формула» графика: ~x1+x2+x3… Помимо указанных функций, код которых генерирует оболочка R commander, существует функция `pairs()`, которая также создаёт графики разброса.

Пример матрицы парных графиков гематологических показателей приведён на рис. 1. По диагонали представлены имена анализируемых переменных. Каждая точка отражает одно наблюдение, её координаты определяются значениями двух переменных. Выше и ниже диагонали с именами переменных расположены одни и те же пары переменных, но по разным осям.

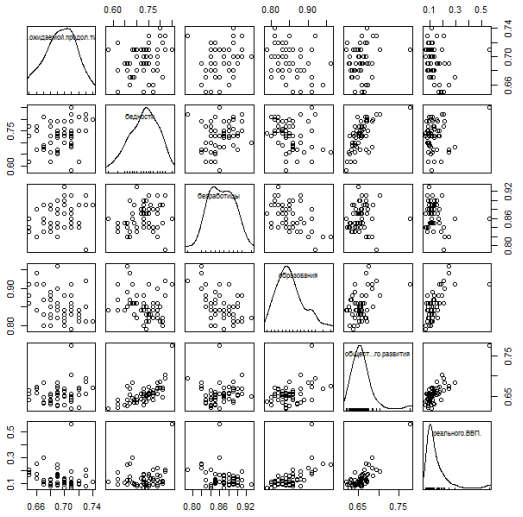

Рисунок 1. Диаграммы рассеяния

Если переменные тесно и линейно связаны, то множество точек данных принимает форму узкого эллипса или почти прямой.

Диаграммы рассеяния предоставляют исследователю больше информации, чем простое значение коэффициента корреляции. Они позволяют:

– выявить отсутствие однородности в выборке (например, наличие подгрупп с разным характером взаимосвязи);

– найти выбросы, или нетипичные данные, которые искусственным образом могут значительно увеличить или уменьшить коэффициент корреляции Пирсона;

– обнаружить нелинейный характер взаимосвязи.

Таким образом, перед проведением корреляционного анализа желательно анализировать графики разброса, с помощью которых можно подобрать оптимальный срез данных для исследования (т.е. выделить определённые подгруппы или, наоборот, объединить разные подгруппы в одну, исключить выскакивающие наблюдения) и применить подходящий вид корреляции (Пирсона или его непараметрических аналогов – Спирмена или Кендалла).

Парная линейная регрессия

Регрессионный анализ решает следующие задачи:

– восстановление зависимости между исследуемыми переменными;

– прогноз зависимой переменной (переменной отклика) по известным независимым переменным (предикторам).

Регрессия бывает:

– по числу предикторов – парная и множественная;

– по форме зависимости – линейная и криволинейная.

Исходные данные для регрессионного анализа представляют собой таблицу (матрицу), в которой строки соответствуют объектам (испытуемым), а столбцы – переменным. Все переменные при этом должны быть измерены в количественной шкале. Одна из переменных определяется исследователем как зависимая, а остальные как независимые переменные.

Парная линейная регрессия в общем случае имеет вид: $y=b_0+b_1x$ Нахождение коэффициентов регрессии основано на методе наименьших квадратов (минимизация суммы квадратов отклонений эмпирических значений признака от теоретических, полученных по уравнению регрессии)

Функции R для построения линейной регрессии

Для построения линейной регрессии в пакете R можно воспользоваться функцией `lm()`:

`lm(formula, data, subset, weights, na.action).`

Аргументы

`formula` – символическое описание восстанавливаемой модели. Для парной линейной регрессии имеет вид $y\sim x$, для множественной – $y\sim x_1+x_2+x_3\ldots$;

`data` – источник данных;

`subset` – подмножество данных, участвующих в построении модели, необязательный параметр;

`weights` – вектор весов, может быть или NULL, или числовым;

`na.action` – обработка пропущенных данных (NA).

Трактовка результатов

Объект, возвращаемый функцией `lm`, имеет различные поля:

`Residuals` – остатки ($y_i–bx_i$), распределение по квартилям.

`Coefficients` – коэффициенты регрессии и их статистическая значимость:

`Estimate` – коэффициент регрессии b;

`Std. Error` – ошибка коэффициента регрессии b;

`t value` – статистика t для оценки уровня значимости коэффициента регрессии;

`Pr(>|t|)` – достигнутый уровень значимости коэффициента регрессии.

`Multiple R-squared` – коэффициент детерминации модели.
`Adjusted R-squared` – скорректированный коэффициент детерминации модели.
`F-statistic` – F-статистика для модели в целом.

Задание и порядок выполнения лабораторной работы №2.3

1. Запустить пакет R commander (Rcmdr). Дальнейшая работа будет проходит в данной графической оболочке, которая генерирует необходимый код через кнопочный интерфейс.

2. Загрузить данные для анализа из файла Данные.xlsx

Для загрузки данных в графической оболочке R commander выбрать следующие пункты меню:

Данные
Импорт данных
Из файла Excel.

3. Исследуйте взаимосвязь заданных индексов. Для этого воспользуйтесь корреляционным анализом:

Статистики
Итоги
Корреляционная матрица
В появившемся диалоговом окне укажите исследуемые переменные.

Выберите вначале корреляцию Пирсона, потом Спирмена. Проанализируйте полученную матрицу корреляций. Какие связи сильные, какие слабые, а какие умеренные? Для нескольких пар показателей найдите уровень значимости коэффициента корреляции.

Статистики
Итоги
Корреляционный тест

4. Проиллюстрируйте полученные результаты (для этих же переменных) на графиках разброса:

Графики
Матрица точечных графиков
Выбрать необходимые переменные.
Убрать галочки с «линии наименьших квадратов» и «сгладить линии».

5. Постройте уравнение зависимости индекса реального ВВП от индекса общественного развития. 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, независимая – индекс общественного развития. Проанализируйте график остатков (наблюдаемое минус предсказанное регрессионной моделью значение). Одинаковы ли они на всём диапазоне значений предсказывающей переменной.

Модели
Графики
График Компонента + остаток

6. По таблице коэффициентов запишите полученное уравнение регрессии

7. Загрузите свои экспериментальные данные. Проведите корреляционный анализ всех данных. Проанализируйте полученную матрицу корреляций. Проиллюстрируйте полученные результаты (для этих же переменных) на графиках разброса.

8. Выберите 3 графика разброса. Укажите численное значение коэффициентов корреляции и их уровней значимости. Дайте содержательную оценку взаимосвязей: прямая или

обратная, слабая или сильная. Какие значения коэффициентов корреляции больше по абсолютному значению: Пирсона или Спирмена?

9. Постройте уравнение зависимости двух переменных. По таблице коэффициентов запишите полученное уравнение регрессии. Проанализируйте график остатков

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 эксперемнтальных данных;
- выводы по п.п.7-9;
- программный код с комментариями;
- выводы по работе.

Контрольные вопросы

1. Корреляционный анализ.
2. Регрессионный анализ.
3. Методы определения корреляционной связи.
4. Коэффициенты корреляции
5. Функции R для проведения корреляционного анализа
6. Функции R для графической интерпретации корреляционного анализа.
7. Функции R для построения линейной регрессии

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.