Министерство науки и высшего образования РФ

ФГБОУ ВО «Ульяновский государственный университет»

Институт экономики и бизнеса

# МЕТОДИЧЕСКИЕ УКАЗАНИЯ И ЗАДАНИЯ К ЛАБОРАТОРНЫМ РАБОТАМ ПО ДИСЦИПЛИНЕ «ВЕРОЯТНОСТНЫЕ МЕТОДЫ В ЭКОНОМИКЕ»

Ульяновск 2018

Методические указания и задания к лабораторным работам по дисциплине «Вероятностные методы в экономике» / составитель: А.Е.Эткин.- Ульяновск: УлГУ, 2018.

Настоящие методические указания предназначены для студентов направлений 38.03.01 «Экономика», 38.03.02 «Менеджмент», 38.03.03 «Управление персоналом», 38.03.05 «Бизнес-информатика», 38.05.01 «Экономическая безопасность». Указания необходимо использовать при выполнении лабораторных работ, предусмотренных учебным планом

Рекомендованы к использованию ученым советом
Института экономики и бизнеса УлГУ
Протокол № 213/09 от 24.05.2018г.

**Лабораторная работа № 1.**
**Статистический подход к определению вероятности.**

Цели работы.

1. Знакомство с методом статистических испытаний.
2. Знакомство с функциями генерации случайных чисел в языках программирования.
3. Получение навыков практического использования метода статистических испытаний для расчета вероятности.
4. Сравнительный анализ различных подходов к определению вероятностей событий.

Постановка задачи.

Существуют различные подходы к определению вероятности: классический, геометрический, аксиоматический, статистический. Каждую из предложенных ниже задач следует решить двумя способами: точным (используя классическое, либо геометрическое определение вероятности) и приближенным (используя статистическое определение), и сравнить результаты между собой. Для статистической оценки вероятности события требуется придумать адекватную модель статистических испытаний, и реализовать ее в MS Excel или с помощью программы, написанной на любом языке программирования.

Задание(список задач).

1. Четыре ботинка выбраны случайно из пяти разных пар. Какова вероятность, что среди них есть хотя бы одна пара?
2. Дверь заперли на два замка, ключи от которых положили в карман, в котором уже лежали 4 ключа. Один из этих 6 ключей потеряли. Какова вероятность того, что
   а) удастся открыть дверь?
   б) подойдут первые два ключа, случайно вынутых из кармана?
3. Автобус должен высадить 15 пассажиров на 4 остановках. Какова вероятность, что
   а) все пассажиры выйдут на одной остановке?
   б) на каждой остановке выйдет хотя бы один человек?
4. Пусть (a, b) - координаты случайной точки в квадрате K= {(a, b): |a|<1, |b|<1}.Найти вероятность того, что корни уравнения $x^2 + ax + b = 0$
   а) действительны;  б) положительны.
5. Расстояние от А до В автобус проходит за 2 минуты, а пешеход за 15 минут. Интервал движения автобусов 25 минут. Пешеход в случайный момент времени отправляется из А в В пешком. Какова вероятность того, что его в пути догонит автобус?

6. *Задача Бюффона*. Игла длины *l* бросается на плоскость, разграфленную параллельными прямыми на полосы шириной *L (L>l)*.Все положения центра иглы и все ее направления равновероятны. Найти вероятность того, что игла пересечет какую-нибудь из линий.

   *Указание*. При статистическом моделировании рассмотреть случай *L=2l*.

Оформление отчета.

Отчет должен содержать для каждой задачи:

1. Аналитическое решение задачи (на основе классического определения вероятности, правил и формул комбинаторики, либо геометрического определения) и расчет точного значения вероятности.
2. Описание модели статистических испытаний для оценки вероятности события.
3. Программу для оценки вероятности события в результате испытаний.
4. Таблицу сравнения оценок вероятности с истинным значением вероятности, в зависимости от числа испытаний. Таблица должна иметь следующий вид:

| | Число испытаний | | | | |
|---|---|---|---|---|---|
| | 10 | 100 | 1000 | 10000 | 100000 |
| Оценка | | | | | |
| Погрешность | | | | | |

Контрольные вопросы для допуска и защиты работы.

1. Сформулируйте классическое определение вероятности. Каковы недостатки этого определения?
2. Сформулируйте геометрическое определение вероятности. Каковы недостатки этого определения? Сопоставьте их с недостатками классического определения.
3. Сформулируйте статистическое определение вероятности. Каковы недостатки этого определения? Сопоставьте их с недостатками классического определения.
4. Существует ли определение вероятности, свободное от указанных выше недостатков? Какое?

Указания по выполнению работы.

Для статистического моделирования рекомендуется использовать языки программирования R или Python, т.к. они имеют функции, наиболее подходящие для

статистического моделирования. Продемонстрируем такое моделирование, используя язык R, на примере следующих задач.

*Задача 1.* Студент выучил 15 из 25 вопросов программы. В билете 4 вопроса. Для сдачи зачета необходимо ответить хотя бы на два из них. Какова вероятность того, что студент сдаст зачет?

*Статистическая модель*. Будем считать, что студент выучил первые 15 вопросов (номера выученных вопросов, очевидно, не будут влиять на результат). Осуществим случайным образом выборку 4 различных целых чисел от 1 до 25. Если по крайней мере 2 из них окажутся не более 15, то результат испытания положительный (зачет сдан). Повторим испытание N раз, и посчитаем количество успешных результатов. Поделив его на N, получим оценку вероятности сдачи зачета.

Ниже приведена программа на языке R, реализующая указанный выше алгоритм.

```
N<-100
s<-0
for (i in 1:N) {
v<- sample(1:25, 4)  # случайная выборка 4-х различных чисел от 1 до 25
s <- s + (sum(v<16)>1)
}
s<-s/N
```

Пояснения (для незнакомых с R):

1) Знак <- означает оператор присваивания (эквивалентен знаку =).

2) *for (i in 1:N)* означает цикл длины N (1:N - целочисленный вектор с компонентами от 1 до N). В фигурных скобках заключено тело цикла.

3) Функция *sample(x, n, replace=F, prob=NULL)* создает вектор, являющийся случайной выборкой *n* компонент вектора *x,* с возвращением или без, в зависимости от параметра *replace* (по умолчанию - без возвращения) и распределением вероятностей, заданным параметром *prob* (по умолчанию - равномерное).

4) Переменная *s* подсчитывает количество успешных результатов испытания.

5) *v<16* - логический вектор, *sum(v<16)* - сумма его компонент, при этом F(FALSE) преобразуется в 0, а T (TRUE) в 1. Таким образом, условие *(sum(v<16)>1)* означает сдачу зачета, и в этом случае значение *s* увеличивается на 1.

Точное значение вероятности может быть рассчитано по формуле $P = m/n$,где

n = C(25, 4) - общее число равновероятных исходов испытания,

m = C(15, 4) + C(15, 3)*10 + C(15, 2)*C(10, 2) – число благоприятных исходов испытания.

Вычисления по этим формулам можно произвести на калькуляторе, в MSExcel, или также в R:

```
m<-choose(15,4)+choose(15,3)*10+choose(15,2)*choose(10,2)
n<-choose(25,4)
m/n
[1] 0.8411067
```

Далее меняем значение N, повторяем вычисления по приведенной выше программе, и составляем таблицу на основании результатов испытаний.

*Задача 2.* Какова вероятность того, что наудачу брошенная в квадрат точка окажется внутри вписанного в него круга?

*Статистическая модель*. Рассмотрим квадрат K = {(x, y): |x|<1, |y|<1}. Осуществим случайным образом выборку двух действительных чисел от -1 до 1 (x и y - координаты случайной точки квадрата). Если $x^2 + y^2 < 1$, то точка лежит внутри вписанного круга. Повторим испытание N раз, и посчитаем количество попаданий в круг. Поделив его на N, получим оценку вероятности попадания точки в круг.

Указанный выше алгоритм легко реализуется в MSExcel: выберем в меню *Данные* пункт *Анализ данных* | *Генерация случайных чисел*.В открывшемся окне диалога выберем *Число переменных*: 2, *Число случайных чисел*: N (конкретное значение), *Распределение*: равномерное, *Параметры: между* -1 и 1.

Точное значение вероятности равно отношению площади круга к площади квадрата: π/4.

Можно реализовать этот алгоритм и в R:

```
N<-100
x<-runif(N, -1, 1)
y<-runif(N, -1, 1)
r<-x*x + y*y
M<-sum(r<1)
M/N
```

Функция runif(n, min=0, max=1) создает вектор из n компонент, равномерно распределенных на отрезке [min, max] (по умолчанию min = 0, max = 1).

**Лабораторная работа № 2.**

**Предельные распределения для биномиального.**

Цели работы.

1. Знакомство с законами распределений случайных величин: биномиальным, пуассоновским, нормальным.
2. Знакомство с функциями распределения и плотностями случайных величин в языках программирования и получение навыков практического использования этих функций для расчета вероятности.
3. Исследование зависимости точности формул Пуассона и Муавра - Лапласа от параметров биномиального распределения.

Постановка задачи.

При большом числе испытаний в схеме Бернулли, вычисления по формуле Бернулли становятся трудоемкими. Для экономии времени (в том числе и компьютерного, т.к. эти расчеты могут быть частью решения более сложной задачи и выполняться многократно) можно использовать приближенные формулы. В зависимости от значения вероятности события, биномиальное распределение приближается к распределению Пуассона или нормальному распределению. Требуется исследовать точность приближения по формулам Пуассона, локальной и интегральной формулам Муавра - Лапласа, в зависимости от значений $n$ и $p$.

Исходные данные.

$N$ - номер варианта (= номеру студента в списке группы).

$a = 0{,}1 \cdot N$.

Случайная величина $X$ распределена по биномиальному закону с параметрами $n$ и $p$.

Каждое задание выполняется для значений $n = 10^k$, где $k = 1, 2, 3, 4$.

Значения вероятности $p$ задаются в каждом задании.

Задание.

1. Принимая $p = \frac{a}{n}$, вычислить вероятность события $P(X \geq 3)$ тремя способами: по формуле Бернулли, по формуле Пуассона, локальной формуле Муавра - Лапласа.

2. Для значения $p = 0{,}1 + 0{,}03 \cdot N$ вычислить вероятность события $P(X = n/2)$ тремя способами: по формуле Бернулли, по формуле Пуассона, локальной формуле Муавра - Лапласа.
3. Для $k_1 = 0{,}2n,\ k_2 = 0{,}7n$ и значений *p* из предыдущего пункта вычислить вероятность события $P(k_1 \le X \le k_2)$ четырьмя способами: по формуле Бернулли, по формуле Пуассона, по интегральной формуле Муавра - Лапласа и по модифицированной интегральной формуле Муавра - Лапласа.
4. Все вычисления провести дважды: в *MS Excel* и *R*.
5. По результатам вычислений для каждой из вероятностей ($P(X \ge 3)$, $P(X = n/2)$, $P(k_1 \le X \le k_2)$) заполнить таблицу следующего вида. Например, для $P(X \ge 3)$:

| n | p | Расчет вероятности $P(X \ge 3)$ по формуле | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Бернулли | Пуассона | | Муавра-Лапласа | | Муавра-Лапласа модиф. | |
| | | | значение | погрешн. | значение | погрешн. | значение | погрешн. |
| 10 | | | | | | | | |
| 100 | | | | | | | | |
| 1000 | | | | | | | | |
| 10000 | | | | | | | | |

6. По результатам расчетов провести анализ и сделать выводы относительно точности приближенных формул.

Оформление отчета.

Отчет должен содержать для каждой из трех вероятностей:

1. Расчетную таблицу в *MS Excel*.
2. Программу расчета вероятностей в *R*.
3. Таблицу с результатами расчета в *R*.
4. Выводы о точности приближенных формул.

Контрольные вопросы для допуска и защиты работы.

1. Дайте определение биномиального распределения. Каковы условия его возникновения?
2. Дайте определение распределения Пуассона. Каковы условия его возникновения?

3. При каком условии биномиальное распределение приближается к распределению Пуассона?
4. При каком условии биномиальное распределение приближается к нормальному?
5. Объясните следующее кажущееся противоречие. С ростом количества испытаний *n* биномиальное распределение приближается к нормальному. В то же время, с ростом *n* биномиальное распределение приближается к распределению Пуассона. Но распределение Пуассона не является нормальным.

Указания по выполнению работы.

В *R* имеются функции для работы с целым рядом распространенных законов распределения вероятностей. В зависимости от назначения, имена этих функций начинаются с одной из следующих четырех букв:

- *d* (от "*density*", плотность): функции плотности вероятности (функция вероятности для дискретных величин);
- *p* (от "*probability*", *вероятность*): функции распределения вероятностей;
- *q* (от "*quantile*", квантиль): функции для нахождения квантилей того или иного распределения;
- *r* (от "*random*", случайный): функции для генерации случайных чисел в соответствии с параметрами того или иного закона распределения вероятностей.

Требующиеся нам функции имеют названия *binom* (биномиальное распределение), *pois* (распределение Пуассона), *norm* (нормальное распределение) с соответствующими приставками *d* или *p*. Например, если нужно вычислить $P(X = 5)$ для случайной величины *X*, распределенной по биномиальному закону, то мы используем функцию *dbinom*, а если нужно вычислить $P(X \leq 5)$, то используем функцию *pbinom*. Для получения справки по любой функции (перечень аргументов, их смысл, значения по умолчанию, примеры использования) можно набрать в консоли ?*имя_функции* и нажать клавишу ENTER. Например, набрав в консоли `?dbinom,` мы увидим в окне справки: *dbinom(x, size, prob, log = FALSE), pbinom(q, size, prob, lower.tail = TRUE, log.p = FALSE).* Здесь *size* - число испытаний, *x* - значение случайной величины, вероятность которого требуется найти (число успехов), *prob* - вероятность появления события в одном испытании. Аргумент *log* (*log.p*) представляет собой логическое значение, которое указывает нужно ли логарифмировать полученные значения вероятностей (по умолчанию он принимает значение *FALSE*, т.е. логарифмировать не нужно). Аргумент *q* указывает значение случайной величины, для которого требуется найти значение функции распределения.

Аргумент *lower.tail* также принимает логические значения и в случае значения *TRUE* (заданного по умолчанию) функция *pbinom* дает вероятность $P(X \leq q)$, т.е. значение функции распределения, а в случае *lower.tail = FALSE* - вероятность противоположного события: $P(X > q)$. Например, для вычисления указанных выше вероятностей $P(X = 5)$ и $P(X \leq 5)$ при *n = 100* и *p = 0.01* нужно, соответственно, ввести *dbinom(x=5, size=100, prob=0.01)* и *pbinom(q=5, size=100, prob=0.01)*. Отметим, что значения аргументов, принятые по умолчанию, можно не указывать, если они совпадают с требуемыми нам. Наименования аргументов также можно не указывать, если аргументы указаны в том же порядке, в котором они следуют в описании функции, т.е. можно ввести соответственно *dbinom(5,100,0.01)* и *pbinom(5,100,0.01)*.

Поскольку все переменные в R предполагаются векторами, можно в качестве аргументов задавать векторы. Например, если мы хотим посчитать значения $P(X = 5)$ при *n = 10, 100, 1000, 10000* и *p=a/n*, мы можем сделать это в R следующим образом:

```
a<-3 #вводим значение a, соответствующее номеру варианта
n<-c(10,100,1000,10000)
p<-a/n
dbinom(5,n,p)
[1] 0.1029193 0.1013081 0.1008691 0.1008239
```

Соответствующие функции в *MS Excel* можно найти в *Мастере функций* в категории *Статистические*: БИНОМРАСП, ПУАССОН, НОРМРАСП. Последний аргумент каждой из этих функций называется "*интегральная"* и представляет собой логическое значение: значение ИСТИНА (можно вводить 1) означает, что рассматривается функция распределения, а значение ЛОЖЬ (можно вводить 0) - функция вероятности (плотности вероятности).

Согласно локальной теореме Муавра-Лапласа, для случайной величины *X*, распределенной по биномиальному закону, имеет место приближенная формула:

$P(X = k) = C_n^k p^k q^{n-k} \approx \frac{e^{-x^2/2}}{\sqrt{2\pi npq}}$, где $x = \frac{k-np}{\sqrt{npq}}$.

Для нахождения вероятности $P(k_1 \leq X \leq k_2)$ используется интегральная формула Муавра-Лапласа:

$P(k_1 \leq X \leq k_2) \approx \Phi(x_2) - \Phi(x_1)$ , где $x_1 = \frac{k_1-np}{\sqrt{npq}}$ , $x_2 = \frac{k_2-np}{\sqrt{npq}}$,

Ф(*x*) - функция Лапласа (функция стандартного нормального распределения *N(0, 1)*).

Лучшее приближение дает модифицированная интегральная формула Муавра-Лапласа:

$P(k_1 \leq X \leq k_2) \approx \Phi(x_2 + \frac{1}{2\sqrt{npq}}) - \Phi(x_1 - \frac{1}{2\sqrt{npq}})$.

**Лабораторная работа № 3.**

**Первичная обработка статистических данных.**

Цели работы.

1. Знакомство с основными понятиями и средствами описания эмпирических данных: выборка, вариационный ряд, статистический ряд, гистограмма, полигон частот, эмпирическая функция распределения, а также выборочными числовыми характеристиками.
2. Знакомство с возможностями визуализации статистических данных в языках программирования и в табличном процессоре MS Excel.
3. Знакомство с функциями для расчета выборочных числовых характеристик.
4. Получение навыков практического применения программных средств для визуализации и простейшей статистической обработки результатов наблюдений (расчета описательных статистик).

Постановка задачи.

При решении практических задач статистического анализа часто приходится сталкиваться с большими массивами исходной информации. Первым этапом их обработки служит обычно их визуализация (графическое представление) и вычисление основных выборочных числовых характеристик, называемых описательными статистиками.

Исходные данные.

Ввиду отсутствия реальных эмпирических данных мы будем использовать для обработки данные, сгенерированные с помощью имеющихся программных средств.

В дальнейшем в формулировках заданий N будет означать номер варианта, совпадающий с номером студента в списке группы.

Задание.

1. Сгенерировать выборку объема $n = 100$ дискретной случайной величины X, принимающей с равными вероятностями целочисленные значения из отрезка [N, N + 19]. Для полученной выборки построить:
   1) вариационный ряд;
   2) статистический ряд;
   3) гистограмму;
   4) полигон частот;

5) график эмпирической функции распределения.
6) найти min, max, размах R, выборочное среднее $\bar{x}$, медиану $\widehat{Me}$, моду $\widehat{Mo}$, дисперсию, стандартное отклонение, стандартную ошибку, коэффициент асимметрии, эксцесс.

2. Сгенерировать выборку объема *n = 300* непрерывной случайной величины X, распределенной по закону, выбираемому из следующей таблицы в соответствии с номером варианта N.

| N | Распределение | Параметры |
|---|---|---|
| 1 | Нормальное | $m = 1, \quad \sigma = 3$ |
| 2 | Показательное | $\lambda = 3$ |
| 3 | Логнормальное | $m = 1, \quad \sigma = 3$ |
| 4 | Пирсона (Хи-квадрат) | *n = 5* |
| 5 | Стьюдента | *n = 5* |
| 6 | F-распределение | $n_1 = 1, \quad n_2 = 10$ |
| 7 | Гамма-распределение | $\lambda = 3, \quad k = 2$ |
| 8 | Бета-распределение | $\alpha = 2, \quad \beta = 3$ |
| 9 | Нормальное | $m = 3, \quad \sigma = 1$ |
| 10 | Показательное | $\lambda = 2$ |
| 11 | Логнормальное | $m = 3, \quad \sigma = 1$ |
| 12 | Пирсона (Хи-квадрат) | *n = 7* |
| 13 | Стьюдента | *n = 7* |
| 14 | F-распределение | $n_1 = 3, \quad n_2 = 5$ |
| 15 | Гамма-распределение | $\lambda = 2, \quad k = 3$ |
| 16 | Бета-распределение | $\alpha = 3, \quad \beta = 2$ |
| 17 | Нормальное | $m = 2, \quad \sigma = 2$ |
| 18 | Показательное | $\lambda = 4$ |
| 19 | Логнормальное | $m = 2, \quad \sigma = 2$ |
| 20 | Пирсона (Хи-квадрат) | *n = 15* |
| 21 | Стьюдента | *n = 15* |
| 22 | F-распределение | $n_1 = 5, \quad n_2 = 15$ |
| 23 | Гамма-распределение | $\lambda = 1, \quad k = 2$ |
| 24 | Бета-распределение | $\alpha = 2, \quad \beta = 2$ |

Для полученной выборки построить:

1) группированный статистический ряд;
2) гистограмму;
3) полигон частот;
4) график эмпирической функции распределения;
5) найти min, max, размах $r$, выборочное среднее $\bar{x}$, медиану $\widehat{Me}$, моду $\widehat{Mo}$, дисперсию, стандартное отклонение, стандартную ошибку, коэффициент асимметрии, эксцесс;
6) рассчитать теоретические значения числовых характеристик, указанных выше, и сравнить их с выборочными.

3. Каждое из заданий выполнить в MS Excel и в R.

Оформление отчета.

Отчет должен содержать по каждому из пунктов задания:

1. Сгенерированную выборку.
2. Статистический ряд.
3. Гистограмму и полигон относительных частот с наложенным на них графиком теоретической функции плотности.
4. График эмпирической функции распределения с наложенным графиком теоретической функции распределения.
5. Таблицу с эмпирическими и теоретическими значениями числовых характеристик.
6. Программу в R для генерации выборки, построения графиков и расчета эмпирических числовых характеристик.
7. Формулы для расчета теоретических значений числовых характеристик.
8. Расчетную таблицу в MS Excel для подготовки графиков и расчета числовых характеристик.

Контрольные вопросы для допуска и защиты работы.

1. Дайте определение генеральной совокупности, выборки, вариационного ряда, статистического ряда, группированного статистического ряда, гистограммы, полигона частот, эмпирической функции распределения.
2. Запишите формулы для вычисления следующих числовых характеристик выборки: выборочное среднее, мода, медиана, дисперсия, исправленная дисперсия, стандартная ошибка.
3. К какой функции приближается гистограмма относительных частот при увеличении объема выборки? Как это обосновать?

4. К какой функции приближается эмпирическая функция распределения при увеличении объема выборки? Как это обосновать?

Указания по выполнению работы.

Для выполнения задания в MS Excel рекомендуется следующий порядок действий.

1. Сгенерировать выборку с требуемым законом распределения. Для этого можно сначала сгенерировать выборку из равномерного распределения на отрезке [0, 1]: выбрать в меню *Данные* | *Анализ данных* | *Генерация случайных чисел* и в открывшемся диалоговом окне указать *Число переменных* - 1, *Число случайных чисел* - объем выборки, *Распределение* - равномерное, *Между* - 0 и 1. Затем нужно к каждому элементу полученной выборки применить функцию, обратную к требуемому распределению: например, для получения логнормального распределения применяем функцию ЛОГНОРМОБР, для распределения Пирсона - ХИ2ОБР и т.п. Все эти функции (за исключением функции показательного распределения, для которой обратная функция легко вычисляется аналитически) находятся в категории *Статистические*. В качестве значения аргумента *Интегральная* следует вводить значение 1 (ИСТИНА), т.к. нас интересует функция распределения, а не плотность. Выборку из нормального распределения можно получить сразу (без дополнительного обратного преобразования), т.к. в списке генерируемых распределений имеется нормальное распределение. Для генерации целочисленной выборки (п.1) можно сгенерировать равномерное распределение на отрезке [N, N + 20] и затем применить функцию округления вниз (ОКРУГЛВНИЗ в категории *Математические*).

2. Для получения вариационного ряда примените к сгенерированной выборке сортировку (упорядочите выборку по возрастанию). Выпишете (зафиксируйте в некоторых ячейках) минимальный ($x_{min}$) и максимальный ($x_{max}$) элементы выборки. Вычислите размах выборки: $r = x_{max} - x_{min}$.

3. Постройте (группированный) статистический ряд, разбив все значения выборки на $k = 20$ групп. В случае целочисленной выборки каждая группа будет соответствовать одному целому числу, и нужно будет посчитать сколько раз оно встретилось в выборке. Для нецелочисленной выборки (п.2) рассчитайте границы интервалов группировки по формулам: $x_0 = x_{min},\ x_i = x_{i-1} + r/k,\ i = 1, 2, \dots, k$. Расположите эти значения в некоторой строке и в следующей строке введите логические формулы для проверки принадлежности элемента выборки соответствующему промежутку. При этом нужно записать имена ячеек (используя знак $) так, чтобы они правильно настраивались при копировании. Например, если выборка содержится в столбце А, а границы интервалов - в

ячейках C5:W5, то в ячейку D6 нужно ввести формулу =ЕСЛИ(И($A1<E$5;$A1>D$5);1;0) и скопировать ее в ячейки E6:V6. В ячейке C6 формулу нужно немного изменить: убрать проверку ограничения снизу, т.е. ввести формулу =ЕСЛИ($A1<D$5;1;0). Если теперь эту строку формул скопировать вниз еще на 299 строк, то мы получим матрицу размера 300x20 из 0 и 1, где 1, стоящая в строке с номером $i$ и столбце с номером $j$ означает, что $i$-й элемент выборки принадлежит интервалу $(x_j, x_{j-1}),\ j = 0,1, \dots, k-1$. Просуммировав теперь значения по столбцам, мы получим количество элементов выборки, попавших в каждый из интервалов группировки. Для получения группированного статистического ряда осталось скопировать границы интервалов и полученные значения частот в два соседних ряда, например, два столбца на другом листе. При этом следует копировать именно значения, а не формулы, используя *специальную вставку*.

4. Подготовьте данные для построения гистограммы, полигона частот и графика теоретической плотности распределения. Пусть в столбце A записаны границы интервалов группировки. Рассчитайте в столбце B середины этих интервалов. В следующих двух столбцах поместите значения относительных частот попадания в соответствующий интервал (напротив середин интервалов). В столбце E рассчитайте значение плотности распределения в соответствующей точке (середине интервала), используя статистическую функцию рассматриваемого распределения.

5. Для построения графиков выделите последние три столбца с данными (построенные на предыдущем шаге) и выберите пункт меню *Вставка* | *Гистограмма* и выберите *Гистограмма с группировкой*. Появится гистограмма с тремя рядами данных. После этого щелкните правой кнопкой мыши на втором столбике (второй ряд данных) и в появившемся контекстном меню выберите пункт *Изменить тип диаграммы для ряда…* В появившемся после этого окне *Изменение типа диаграммы* выберите тип *График* и нажмите ОК. После этого график для второго ряда данных будет представлен полигоном частот. Теперь щелкните правой кнопкой мыши на третьем столбике (третий ряд данных) и в появившемся контекстном меню снова выберите пункт *Изменить тип диаграммы для ряда…,* но на этот раз выберите тип *Точечная с гладкими кривыми*. После нажатия на кнопку ОК получим представление третьего ряда в виде кривой - теоретической функции плотности.

6. Подготовьте данные для построения эмпирической и теоретической функций распределения. Для этого скопируйте полученный ранее вариационный ряд на новый лист (в столбец A). В следующем столбце нужно ввести накопленные относительные частоты. Они представляют собой фактически номер соответствующего элемента вариационного ряда, деленный на объем выборки n. Для этого можно, например, в столбец D ввести

последовательно числа от 1 до n, а в ячейку B1 ввести формулу =D1/n (вместо n, естественно, должно быть написано соответствующее значение или указан абсолютный адрес ячейки, в которой оно содержится) и скопировать ее на весь столбец. В столбце C рассчитайте значение теоретической функции распределения (интегральной!) в соответствующей точке, используя статистическую функцию рассматриваемого распределения.

7. Для построения графиков эмпирической и теоретической функций распределения выделите столбцы A, B и C и выберите пункт меню *Вставка Диаграммы| Точечная с гладкими кривыми.* После нажатия на клавишу ОК появится окно с требуемыми графиками.

8. Для расчета описательных статистик составьте таблицу из двух строк: в первой строке введите наименования (обозначения) статистик, а под ними - соответствующие функции/формулы для их вычисления. Все необходимые функции можно найти в категории *Статистические*. Для быстрого расчета всех описательных статистик можно использовать в меню *Данные* надстройку *Анализ данных | Описательная статистика*. В открывшемся диалоговом окне нужно отметить входной интервал данных и поставить флажок *Итоговая статистика*. Отметим, что в выходных данных *интервал* означает размах выборки.

Рассмотрим теперь порядок выполнения задания в RStudio.

1. Как уже указывалось в ЛР №2, для генерации в R выборки с заданным законом распределения используются функции, начинающиеся с буквы *r*, с последующим названием распределения, т.е. *rnorm, rexp, rlnorm, rchisq, rt, rf, rgamma, rbeta*. Каждая из этих функций имеет в качестве первого аргумента объем выборки, далее следуют параметры распределения.

2. Для построения гистограммы и полигона частот в R нет необходимости строить предварительно статистический ряд. В R есть библиотека *ggplot2*, которая содержит все функции, необходимые для их построения. Если пакет *ggplot2* не установлен, то предварительно его нужно установить командой *install.packages(' ggplot2').*
Рассмотрим, в качестве примера, выполнение задания в R для показательного распределения.

```
library(ggplot2) #загружаем библиотеку
n<-300 #вводим объем выборки
k<-20  #число интервалов группировки
x<-rexp(n,5) #генерируем выборку
y<-dexp(x,5) #вычисляем вектор соответствующих значений плотности вероятности
df<-data.frame(x,y) #составляем таблицу из двух столбцов
g <- ggplot(df, aes(x=x)) #формируем разметку области построения графика
```

```
g <- g + geom_histogram(aes(y=..density..), bins=k, alpha=0, color='black')
        #добавляем гистограмму относительных частот
g <- g + geom_freqpoly(aes(y=..density..), bins=k, color='blue')
        #добавляем полигон относительных частот
g <- g + geom_line(aes(y=y), color='red')
        #добавляем график теоретической плотности
g       #выводим график на экран
```

3. Построение эмпирической и теоретической функций распределения можно осуществить следующим образом.

```
df$y<-pexp(x, 5) #заменяем столбец y на значения функции распределения
g <- ggplot(df, aes(x=x)) #формируем разметку области построения графика
g <- g + stat_ecdf() #добавляем график эмпирической функции распределения
g <- g + geom_line(aes(y=y), color='red')
        #добавляем график теоретической функции распределения
g       #выводим график на экран
```

4. В R также имеются встроенные функции для вычисления большинства описательных статистик: *min, max, mean* - выборочное среднее, *median, var* - дисперсия, *sd* - стандартное отклонение. Можно вывести сразу несколько показателей с помощью функции *summary*: минимум, максимум, среднее, медиану, квартили. Для вывода всех требуемых показателей нужно загрузить командой *library(psych)* библиотеку *psych*, которая используется для психометрических исследований, и затем использовать функцию *describe* из этого пакета. Отметим, что в выходных данных *vars* означает количество переменных (не путать с обозначением *var* для дисперсии), *range* - размах, *skew* - асимметрия, *kurtosis* - эксцесс.

**Лабораторная работа № 4.**

**Оценка линейных регрессионных моделей.**

Цели работы.

1. Знакомство с *мастером диаграмм* в *MS Excel* и его практическим использованием для наглядного представления и анализа данных.
2. Знакомство с инструментами графического представления данных в языке R.
3. Изучение и получение навыков практического использования встроенных статистических функций в *MS Excel*.
4. Знакомство с *Пакетом анализа* в *MS Excel* и его использованием для анализа данных.
5. Получение навыков практического использования функций R для анализа данных.

Исходные данные.

В R имеется большое количество встроенных датасетов. В этой работе используется один из них: mtcars. Соответствующий файл содержит данные, взятые из американского журнала Motor Trend 1974 года, о расходе топлива и 10 аспектах дизайна и производительности для 32 автомобилей (модели 1973-74 годов).

| Марки автомобилей | mpg | disp | hp | drat | wt | qsec |
|---|---|---|---|---|---|---|
| Mazda RX4 | 21 | 160 | 110 | 3,9 | 2,62 | 16,46 |
| Mazda RX4 Wag | 21 | 160 | 110 | 3,9 | 2,875 | 17,02 |
| Datsun 710 | 22,8 | 108 | 93 | 3,85 | 2,32 | 18,61 |
| Hornet 4 Drive | 21,4 | 258 | 110 | 3,08 | 3,215 | 19,44 |
| Hornet Sportabout | 18,7 | 360 | 175 | 3,15 | 3,44 | 17,02 |
| Valiant | 18,1 | 225 | 105 | 2,76 | 3,46 | 20,22 |
| Duster 360 | 14,3 | 360 | 245 | 3,21 | 3,57 | 15,84 |
| Merc 240D | 24,4 | 146,7 | 62 | 3,69 | 3,19 | 20 |
| Merc 230 | 22,8 | 140,8 | 95 | 3,92 | 3,15 | 22,9 |
| Merc 280 | 19,2 | 167,6 | 123 | 3,92 | 3,44 | 18,3 |
| Merc 280C | 17,8 | 167,6 | 123 | 3,92 | 3,44 | 18,9 |
| Merc 450SE | 16,4 | 275,8 | 180 | 3,07 | 4,07 | 17,4 |
| Merc 450SL | 17,3 | 275,8 | 180 | 3,07 | 3,73 | 17,6 |
| Merc 450SLC | 15,2 | 275,8 | 180 | 3,07 | 3,78 | 18 |
| Cadillac Fleetwood | 10,4 | 472 | 205 | 2,93 | 5,25 | 17,98 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Lincoln Continental | 10,4 | 460 | 215 | 3 | 5,424 | 17,82 |
| Chrysler Imperial | 14,7 | 440 | 230 | 3,23 | 5,345 | 17,42 |
| Fiat 128 | 32,4 | 78,7 | 66 | 4,08 | 2,2 | 19,47 |
| Honda Civic | 30,4 | 75,7 | 52 | 4,93 | 1,615 | 18,52 |
| Toyota Corolla | 33,9 | 71,1 | 65 | 4,22 | 1,835 | 19,9 |
| Toyota Corona | 21,5 | 120,1 | 97 | 3,7 | 2,465 | 20,01 |
| Dodge Challenger | 15,5 | 318 | 150 | 2,76 | 3,52 | 16,87 |
| AMC Javelin | 15,2 | 304 | 150 | 3,15 | 3,435 | 17,3 |
| Camaro Z28 | 13,3 | 350 | 245 | 3,73 | 3,84 | 15,41 |
| Pontiac Firebird | 19,2 | 400 | 175 | 3,08 | 3,845 | 17,05 |
| Fiat X1-9 | 27,3 | 79 | 66 | 4,08 | 1,935 | 18,9 |
| Porsche 914-2 | 26 | 120,3 | 91 | 4,43 | 2,14 | 16,7 |
| Lotus Europa | 30,4 | 95,1 | 113 | 3,77 | 1,513 | 16,9 |
| Ford Pantera L | 15,8 | 351 | 264 | 4,22 | 3,17 | 14,5 |
| Ferrari Dino | 19,7 | 145 | 175 | 3,62 | 2,77 | 15,5 |
| Maserati Bora | 15 | 301 | 335 | 3,54 | 3,57 | 14,6 |
| Volvo 142E | 21,4 | 121 | 109 | 4,11 | 2,78 | 18,6 |

Здесь:

mpg (miles per gallon) - расход топлива (миль/галлон),

disp (displacement) - объем двигателя (в куб. дюймах),

hp (horsepower) - мощность двигателя (л.с.),

drat (rear-axle ratio) - передаточное число заднего моста,

wt (weight) - вес (в 1000 фунтов),

qsec (1/4 mile time) - время разгона (в секундах).

Из таблицы удалены столбцы, соответствующие фиктивным (dummy) переменным, т.к. в данной работе они не используются.

Постановка задачи.

Выбрать объясняемые и объясняющие переменные. Используя инструменты наглядного представления данных и их анализа, отобрать переменные, от которых наиболее всего зависят объясняемые переменные. Исследовать корреляционные зависимости между переменными. Построить линейные регрессионные модели для объясняемых переменных.

Задание.

1. Для одной из выбранных объясняемых переменных и одной из объясняющих переменных построить точечную диаграмму в MS Excel.
2. Добавить на построенной диаграмме линию регрессии.
3. Добавить на построенной диаграмме уравнение регрессии и вывести значение коэффициента детерминации.
4. Получить оценку регрессионной зависимости между выбранными переменными, используя *Мастер функций MS Excel*.
5. Вывести график зависимости и линию регрессии для тех же переменных в R.
6. Получить в R оценку линейной модели парной регрессии для выбранных переменных.
7. Построить в R графики всевозможных зависимостей между парами переменных.
8. Вывести в R матрицу парных корреляций между всевозможными парами переменных.
9. Получить в R оценки моделей выбранных объясняемых переменных от остальных переменных.
10. Выполнить задания 8 и 9 в *MS Excel*, используя надстройку *Анализ данных*. Сравнить полученные результаты с результатами, полученными в заданиях 8 и 9.
11. Дать описание полученных моделей: интерпретировать коэффициенты, проверить их значимость и значимость модели в целом, оценить качество модели.
12. Получить оценки тех же объясняемых переменных только от значимых объясняющих переменных. Повторить для них задание п.11. Как изменилось качество моделей?

Оформление отчета.

Отчет должен содержать:

1. Точечную диаграмму MS Excel зависимости между выбранными переменными с указанными на ней уравнением регрессии и коэффициентом детерминации.
2. Оценки модели парной линейной зависимости между выбранными переменными, полученные в MS Excel и в R.
3. Диаграмму рассеяния в R для выбранной пары переменных и для всевозможных пар переменных.
4. Корреляционные матрицы зависимости между всевозможными парами переменных, полученные в R и в MS Excel.
5. Таблицы с решениями задач пп.4, 6, 9, 12 задания.

6. Оценки линейных моделей множественной регрессии объясняемых переменных на все объясняющие переменные, полученные в MS Excel и в R.
7. Описание результатов оценки моделей: значимость коэффициентов, значимость модели в целом, качество модели.
8. Интерпретацию коэффициентов модели.

Контрольные вопросы для допуска и защиты работы.

1. Что такое регрессионная зависимость? Чем она отличается от функциональной?
2. Записать общий вид уравнения линейной регрессии и пояснить обозначения.
3. Каким требованиям должны удовлетворять переменные в регрессионных моделях? Почему?
4. Какие предположения делаются об ошибках в линейных регрессионных моделях?
5. Как связаны коэффициент корреляции и коэффициент детерминации в модели парной линейной регрессии?
6. Что такое коэффициент множественной корреляции и как он связан с коэффициентом детерминации в случае множественной регрессии?
7. По какому закону распределена статистика для проверки гипотезы о значимости коэффициента в уравнении регрессии?
8. По какому закону распределена статистика для проверки гипотезы о значимости регрессии в целом?

Указания по выполнению работы.

1. Для работы в *MS Excel* скопировать исходные данные на лист *MS Excel.*
2. Выделите мышкой столбцы, соответствующие выбранным переменным. Для построения диаграммы выбрать в главном меню пункт *Вставка* и из меню *Диаграммы* выбрать вид: *точечная с маркерами.*
3. Щелкнув правой кнопкой мыши на любой из точек диаграммы, выбрать в открывшемся контекстном меню: *Добавить линию тренда*, и в открывшемся диалоговом окне выбрать вид линии: *Линейная*, и поставить флажки, соответствующие отображению на диаграмме уравнения линии и коэффициента детерминации.
4. Для оценки регрессионной зависимости с помощью *Мастера функций* в *MS Excel* нужно последовательно выполнить следующие шаги:
    1) Выделить (пометить мышкой) область пустых ячеек размером 5 x 2 (5 строк, 2 столбца).

2) Вызвать *Мастер функций* (нажать на кнопку $f_x$ в строке формул), в открывшемся диалоговом окне выбрать категорию функций: *Статистические*, и среди них выбрать функцию ЛИНЕЙН.
3) В открывшемся окне диалога нужно ввести аргументы функции: *y* - значения объясняемой переменной, *x* - значения объясняющей переменной (соответствующие столбцы помечаем мышкой, и адреса ячеек вводятся в нужные поля). Аргумент *Константа* - логическое значение, указывающее на наличие свободного члена в уравнении регрессии (обычно вводим 1, что соответствует значению ИСТИНА). Последний аргумент *Статистика* - это также логическое значение, указывающее на необходимость вывода дополнительной информации по статистике (обычно вводим 1, и получаем всю информацию, если же ввести 0, т.е. ЛОЖЬ, то получим только коэффициенты регрессии).
4) После ввода аргументов функции и нажатия клавиши ОК, в левой верхней ячейке выделенной области появится первое значение: коэффициент наклона в регрессии. После этого нужно нажать клавишу F2, а затем одновременно нажать 3 клавиши: CTRL+SHIFT+ENTER, после чего вся выделенная область заполнится значениями в следующем порядке:

| | |
|---|---|
| $\hat{b}$ | $\hat{a}$ |
| $s_{\hat{b}}$ | $s_{\hat{a}}$ |
| $R^2$ | $s$ |
| $F$ | $n-2$ |
| $ESS$ | $USS$ |

5. Для работы в R будем использовать RStudio. Для вывода графика зависимости в R можно использовать функцию *plot(x, y, …).* Например, для вывода графика зависимости переменной *mpg* от переменной *disp*, достаточно ввести в окне редактора скриптов, либо в окне консоли *plot(mtcars$disp, mtcars$mpg, xlab = 'disp', ylab = 'mpg'),* после чего, соответственно, нажать на кнопку *Run* или клавишу *Enter*.
6. Для оценки модели парной линейной регрессии в R можно использовать функцию *lm(formula, data)*, где *formula (*формула модели*)* представляется в виде *y ~ x1 + x2 + ...*, где *y* - объясняемая переменная, *x1, x2, ...* - объясняющие переменные. Например, для оценки зависимости переменной *mpg* от переменной *disp,* применим

функцию *lm* и сохраним результаты оценки в переменной *model*. Для просмотра результатов оценки модели используем функцию *summary*:

```
model<-lm(mpg~disp, mtcars)
        summary.lm(model)
```

Для добавления регрессионной прямой к графику, построенному выше (п.5), можно использовать функцию *abline:*

```
         abline(model)
```

7. Для построения графиков всевозможных пар переменных в R можно использовать функцию *pairs*. Предварительно следует просмотреть структуру датасета с помощью функции *str*, и сделать отбор переменных:

```
str(mtcars)
df<-mtcars[c(1,3,4,5,6,7)] #отбираем столбцы с указанными номерами
pairs(df)
```

8. Для оценки корреляционной матрицы в R можно воспользоваться функцией *cor*. Для той же цели в *MS Excel* можно использовать надстройку *Анализ данных* из меню *Данные*, в которой нужно выбрать программу *Корреляция*. В открывшемся окне диалога рекомендуется поставить флажок *Метки в первой строке* и в качестве входного интервала пометить всю таблицу данных вместе с названиями столбцов.
9. Оценка модели линейной регрессии в MS Excel также может быть осуществлена с помощью надстройки *Анализ данных* | *Регрессия*. В окне диалога также разумно поставить флажок *Метки* и, соответственно, помечать данные вместе с названиями. При этом в качестве входного интервала Y помечается один столбец, соответствующий объясняемой переменной, а в качестве входного интервала X - все столбцы объясняющих переменных (вся матрица регрессоров). Флажок *Константа-ноль* ставится только в том случае, когда свободный член в уравнении регрессии заведомо равен нулю. Если нужны доверительные интервалы (для коэффициентов регрессии) для доверительной вероятности, отличной от 0,95 (принимаемой по умолчанию), то следует поставить флажок *Уровень надежности,* и указать соответствующую доверительную вероятность. Доверительные интервалы для $\delta = 0{,}95$ будут выведены в любом случае. В параметрах вывода можно оставить *Новый рабочий лист* (результаты оценки регрессии будут представлены на другом листе). Остальные настройки нам в этой работе не понадобятся. Результаты оценки регрессии представляются в виде трех таблиц (см. пример ниже).

ВЫВОД ИТОГОВ

| *Регрессионная статистика* | |
|---|---|
| Множественный R | 0,872294085 |
| R-квадрат | 0,76089697 |
| Нормированный R-квадрат | 0,744407106 |
| Стандартная ошибка | 3,046995662 |
| Наблюдения | 32 |

Дисперсионный анализ

| | *df* | *SS* | *MS* | *F* | *Значимость F* |
|---|---|---|---|---|---|
| Регрессия | 2 | 856,8058932 | 428,4029466 | 46,14331351 | 9,76076E-10 |
| Остаток | 29 | 269,2412943 | 9,284182562 | | |
| Итого | 31 | 1126,047188 | | | |

| | *Коэффициенты* | *Стандартная ошибка* | *t-статистика* | *P-Значение* | *Нижние 95%* | *Верхние 95%* | *Нижние 99,0%* | *Верхние 99,0%* |
|---|---|---|---|---|---|---|---|---|
| Y-пересечение | 30,29037038 | 7,317878326 | 4,13922848 | 0,000273782 | 15,32362894 | 45,25711183 | 10,11947373 | 50,46126704 |
| drat | 1,442490742 | 1,458567604 | 0,988977637 | 0,33085441 | -1,540614911 | 4,425596394 | -2,57788443 | 5,462865922 |
| wt | -4,782890199 | 0,79703526 | -6,000851451 | 1,58907E-06 | -6,413010314 | -3,1527701 | -6,97982695 | -2,58595344 |

Дадим пояснения к этим таблицам (там, где это требуется). В первой таблице (Регрессионная статистика): *Множественный R* - множественный коэффициент корреляции, *Нормированный R-квадрат* - скорректированный коэффициент детерминации. Во второй таблице (Дисперсионный анализ): *df* - числа степеней свободы (*m-1, n-m* и *n-1* соответственно, где *n* - число наблюдений, *m* - число параметров), *SS* - суммы квадратов (*ESS, USS* и *TSS* соответственно). *MS* - среднее значение суммы квадратов в расчете на одну степень свободы (*SS/df*). Значимость F - вероятность таких результатов (наблюдений) при условии, что регрессия не значима (т.е. нет зависимости объясняемой переменной от рассматриваемых объясняющих переменных). Заметим, что это значение часто указывается в экспоненциальной форме (9,76076E-10 означает $9{,}76076*10^{-10}$). В последней таблице во втором столбце указаны оценки коэффициентов регрессии при переменных, указанных в первом столбце (Y-пересечение соответствует свободному члену). P-значение - вероятность того, что соответствующий коэффициент не значим. Далее следуют две пары столбцов, содержащие, соответственно, нижние и верхние границы доверительных интервалов для указанных доверительных вероятностей.

**Лабораторная работа № 5.**

**Нелинейные модели регрессии и их линеаризация.**

Цели работы.

1. Знакомство с нелинейными моделями регрессии и методами их оценки.
2. Получение навыков практического применения метода линеаризации к нелинейным моделям.
3. Получение навыков практической проверки предположений, лежащих в основе классической модели регрессии.
4. Знакомство с методами верификации модели.

Исходные данные.

Исходные данные те же, что и в работе №4.

Постановка задачи.

Наглядное графическое представление зависимостей между переменными из датасета *mtcars* (с использованием мастера диаграмм в *MS Excel* или функции *plot* в *RStudio*) показывает, что эти зависимости в основном нелинейные. Поэтому можно предположить, что более адекватными моделями будут нелинейные модели регрессии.

Поэтому предлагается построить нелинейную модель, линеаризовать ее, оценить линеаризованную модель и сравнить новую модель с соответствующей линейной моделью. Выберите переменную *mpg* (расход топлива) в качестве объясняемой переменной, а переменные *wt* (вес) и *hp* (мощность двигателя) в качестве объясняющих. Рассмотрите в качестве спецификации модели степенную зависимость переменных. Линеаризуйте эту модель с помощью операции логарифмирования. Сравните полученную модель с соответствующей линейной моделью. Поскольку сравнение различных видов зависимостей по коэффициенту детерминации (или любому другому показателю, вычисляемому при оценке регрессии) некорректно, то для сравнения следует построить модели по половине данных, а по второй половине провести их сравнение.

Задание.

1. Выведите графики зависимостей переменной *mpg* от каждой из переменных *wt* и *hp,* добавив подходящие линии тренда.

2. Прологарифмируйте значения переменных и постройте графики зависимостей *ln(mpg)* от *ln(wt)* и от *ln(hp)*. Похожа ли теперь зависимость на линейную? Добавьте линии трендов.
3. Оцените модели линейной и степенной зависимости *mpg* от совокупности переменных *wt* и *hp*. В какой из моделей выше коэффициент детерминации?
4. Выведите для каждой из моделей график остатков и проверьте его визуально на соответствие основным предположениям классической линейной регрессионной модели. Какие из предположений, на ваш взгляд, выполняются (не выполняются) для каждой из моделей. Какая из моделей является более адекватной?
5. Выберите случайным образом из таблицы данных *mtcars* половину наблюдений. Проведите оценку линейной и степенной модели по выбранной половине наблюдений.
6. Осуществите прогноз значений *mpg* для второй половины наблюдений, используя каждую из моделей. Рассчитайте сумму квадратов отклонений истинных значений от предсказанных для каждой из моделей. Какая модель лучше?
7. Выполните задания 1-6 в *MS Excel* и в *RStudio*.

Оформление отчета.

Отчет должен содержать:

1. Графики в *Мастере диаграмм MS Excel* зависимостей переменной *mpg* от каждой из переменных *wt* и *hp,* и соответствующие линии тренда, их уравнения и коэффициенты детерминации.
2. Графики в *MS Excel* и в *RStudio* зависимостей *ln(mpg)* от *ln(wt)* и от *ln(hp)* с добавлением линий трендов. Текст программы в R для вывода графиков.
3. Таблицы в *MS Excel* с результатами оценок моделей линейной и степенной зависимости *mpg* от совокупности переменных *wt* и *hp*, включая вывод графиков остатков.
4. Текст программы в R для оценки моделей линейной и степенной зависимости *mpg* от совокупности переменных *wt* и *hp,* и отображения результатов. Таблицы в *RStudio* с результатами оценок моделей.
5. Вывод графиков остатков в *RStudio* для каждой из построенных моделей. Текст соответствующей программы в R для расчета и вывода остатков.
6. Текст программы в R для выбора половины наблюдений, оценки моделей и расчета суммы квадратов отклонений истинных значений от предсказанных для второй половины наблюдений по каждой из моделей.

7. Расчетные таблицы в MS Excel для оценки моделей по тем же выбранным строкам, и расчета тех же значений, что и в предыдущем пункте.

Контрольные вопросы для допуска и защиты работы.

1. Можно ли к нелинейной модели регрессии применять метод наименьших квадратов?
2. В чем сложность оценки нелинейной модели регрессии?
3. Что такое линеаризация модели? Перечислите основные случаи, когда она возможна.
4. Что является основным источником ошибки в модели регрессии?
5. Каковы основные предположения об ошибках в модели регрессии?
6. Имеется две модели регрессии. Как определить какая из моделей лучше?

Указания по выполнению работы.

1. Для вывода графика остатков при оценке модели в *MS Excel* нужно в диалоговом окне *Регрессия* поставить флажок *График остатков*, а для проверки распределения остатков на нормальность - поставить флажок *График нормальной вероятности.*
2. Для вывода графика остатков и его последующей проверки на соответствие предположениям классической регрессионной модели в R можно использовать функцию *plot*, применив ее к результату оценки модели. Например, если *model* есть модель, полученная в результате применения функции *lm*, то введя команду

```
plot(model)
```

получим сообщение

```
Hit <Return> to see next plot:
```

Далее, нажимая последовательно на клавишу *Enter*, получим 4 графика. Первый график дает распределение остатков в зависимости от предсказанных значений объясняемой переменной. Второй график дает зависимость стандартизованных остатков от теоретических квантилей нормального распределения. Чем лучше точки ложатся на пунктирную прямую, тем ближе распределение остатков к нормальному. На третьем графике представлена зависимость стандартизованных остатков от предсказанных значений объясняемой переменной. По этому графику можно оценить выполнение предположения о гомоскедастичности ошибок регрессии. Если на графике нет выраженной зависимости остатков от предсказанных значений объясняемой переменной, то можно считать, что предположение о гомоскедастичности ошибок выполняется.

Последний график служит для выявления выбросов - наблюдений, которые плохо предсказываются моделью, построенной по остальным наблюдениям. Удаление этих наблюдений (выбросов) существенно изменяет коэффициенты модели и повышает качество приближения.

3. Для отбора половины строк случайным образом можно воспользоваться функцией *sample*:

```
v<-sample(1:32,16) #отбираем в вектор v 16 случайных чисел от 1 до 32
df1<-mtcars[v, ] #в таблицу df1 записываем строки с отобранными    номерами
df2<-mtcars[-v, ] #в таблицу df2 записываем оставшиеся строки
```

4. Для расчета прогнозируемых значений в R можно использовать функцию *predict(model, args)*, где *model* - построенная модель, а *args* - таблица данных со значениями аргументов. При этом следует помнить, что в случае степенной модели мы получаем прогноз не самой объясняемой переменной, а ее логарифма. Поэтому полученные с помощью функции *predict* значения следует пропотенцировать. Для аналогичных расчетов в *MS Excel* потребуется ввести формулу, соответствующую рассматриваемой модели, с использованием полученных оценок коэффициентов модели.
5. Расчет суммы квадратов отклонений наблюдаемых значений объясняемой переменной от предсказанных с помощью модели можно выполнить в MS Excel с помощью функции *СУММКВРАЗН***,** в качестве аргументов которой указать соответствующие столбцы значений. Аналогичные расчеты в R можно осуществить с помощью конструкции *sum((y-y1)*(y-y1))*, где *y* и *y1* - соответственно векторы наблюдаемых и предсказанных значений. Все операции в R векторизованы, т.е. выполняются над всеми компонентами векторов. Поэтому *(y-y1)*(y-y1)*дает нам вектор квадратов остатков, а функция *sum* подсчитывает затем сумму его компонент.

**Лабораторная работа № 6.**

**Мультиколлинеарность.**

Цели работы.

1. Знакомство с понятием мультиколлинеарности и проблемами, возникающими при оценке регрессионных моделей с мультиколлинеарностью.
2. Получение навыков практической проверки наличия мультиколлинеарности в исходных данных.
3. Знакомство с основными методами устранения или уменьшения мультиколлинеарности.
4. Получение навыков практического применения регрессионных моделей при наличии мультиколлинеарности в исходных данных.

Исходные данные.

В этой работе мы продолжаем моделирование на уже известном наборе данных *mtcars*.

Постановка задачи.

В работе 2 мы уже видели, что более адекватной моделью зависимости между переменными датасета *mtcars* является степенная модель. Поэтому прологарифмируем значения исходных переменных и будем рассматривать модель линейной зависимости *ln(mpg)* от логарифмов остальных переменных. МНК-оценка этой модели показывает, что все переменные в этой модели, кроме одной, незначимы. В то же время модель в целом имеет высокую значимость. Это может быть следствием мультиколлинеарности модели. Предположение о мультиколлинеарности подтверждается корреляционной матрицей и расчетом показателей вздутия дисперсии. Поэтому предлагается применить различные методы устранения или уменьшения мультиколлинеарности для построения наиболее адекватной модели.

Задание.

1. Прологарифмируйте значения переменных *mpg, disp, hp, drat, wt, qsec* из датасета *mtcars* и оцените модель линейной зависимости *ln(mpg)* от логарифмов остальных переменных. Значима ли полученная зависимость? Сколько из коэффициентов при переменных в этой модели оказались значимыми?

2. Выведите корреляционную матрицу для указанных выше переменных. Какие из элементов этой матрицы свидетельствуют о мультиколлинеарности?
3. Рассчитайте показатели вздутия дисперсии для каждой из независимых переменных. Подтверждают ли они наличие мультиколлинеарности?
4. Устраните мультиколлинеарность, используя алгоритм пошагового отбора наиболее информативных переменных. Реализуйте этот алгоритм в MS Excel, используя как версию отбора *вперед*, так и *назад*. Совпали ли полученные модели? Если нет, то выберите ту, которая является лучшей на ваш взгляд.
5. Примените для устранения мультиколлинеарности метод главных компонент. Реализуйте алгоритм этого метода в RStudio. Постройте регрессию объясняемой переменной на все главные компоненты. Какие из них оказались значимыми? Оцените регрессию только от значимых главных компонент.
6. Выведите таблицу весов, с которыми объясняющие переменные входят в каждую из главных компонент. Чему равен коэффициент корреляции первой главной компоненты с объясняемой переменной *ln(mpg).* Какая доля дисперсии объясняемой переменной, приходится на каждую из главных компонент? Постройте соответствующую гистограмму.
7. Сопоставьте модели, полученные пошаговым отбором переменных и методом главных компонент. Какая из них лучше, на ваш взгляд, и почему?

Оформление отчета.

Отчет должен содержать:

1. Оценку модели линейной зависимости *ln(mpg)* от логарифмов остальных переменных.
2. Корреляционную матрицу логарифмов всех переменных.
3. Показатели вздутия дисперсии для всех объясняющих переменных модели.
4. Все промежуточные и итоговые оценки моделей при пошаговом отборе объясняющих переменных.
5. Матрицу весов, с которыми объясняющие переменные входят в каждую из главных компонент.
6. Коэффициент корреляции *ln(mpg)* и первой главной компоненты.
7. Итоговую информацию по распределению дисперсии объясняемой переменной по главным компонентам, представленную в табличном и графическом виде.
8. Оценки регрессии *ln(mpg)* на все главные компоненты и только на значимые главные компоненты.

Контрольные вопросы для допуска и защиты работы.

1. В чём суть явления мультиколлинеарности?
2. Чем отличается строгая (точная) мультиколлинеарность от нестрогой? Какая из них представляет более серьезную проблему? Почему?
3. Как обнаружить мультиколлинеарность?
4. Возможна ли оценка модели при наличии мультиколлинеарности?
5. Можно ли использовать оценку модели с мультиколлинеарностью для прогнозирования?
6. Каковы известные методы устранения/уменьшения мультиколлинеарности? Опишите кратко суть их алгоритмов и условия применения.

Указания по выполнению работы.

1. Оценку модели линейной зависимости *ln(mpg)* от логарифмов остальных переменных можно провести как в *MS Excel*, так и в *R*.
2. Показатель вздутия дисперсии (VIF - Variance Inflation Factor) рассчитывается по формуле $VIF_k = \frac{1}{1-R_k^2}$, где $R_k^2$ - коэффициент детерминации в регрессии k-ой объясняющей переменной на все остальные объясняющие переменные. Для их расчета в *MS Excel* придется оценивать 5 регрессий (столько, сколько объясняющих переменных). В *R* это сделать значительно проще: показатели *VIF* рассчитываются с помощью одноименной функции сразу для всех регрессоров ранее оцененной модели:

```
df<-mtcars[c(1,3,4,5,6,7)] # выбираем в датафрейм нужные столбцы
ln_df<-data.frame(sapply(df, log)) # логарифмируем данные
model<-lm(mpg~disp+hp+drat+wt+qsec, ln_df) # оцениваем модель
vif(model) # рассчитываем показатели VIF
```

   Функция *VIF* находится в пакете *car*. Поэтому нужно предварительно загрузить пакет командой `library('car')`. Если же пакет не установлен, то предварительно его нужно установить командой `install.packages('car')`.
3. При пошаговом отборе переменных *«вперед»,* на первом шаге выбирается объясняющая переменная, имеющая с объясняемой набольший по модулю коэффициент корреляции (выбираем ее на основе выведенной корреляционной матрицы), и оценивается соответствующая модель. Далее рассматриваем всевозможные пары объясняющих переменных, содержащих выбранную на первом шаге, и оцениваем соответствующие модели. Из полученных моделей

отбираем ту, которая имеет наибольший коэффициент детерминации. Если скорректированный коэффициент детерминации для этой модели выше, чем у полученной на предыдущем шаге, то процесс продолжается (рассматриваются тройки объясняющих переменных и т.д.). Если же скорректированный коэффициент детерминации понизился, то останавливаемся на модели, полученной на предыдущем шаге. Аналогично осуществляется пошаговый отбор *«назад».* При этом нужно на первом шаге включить в регрессию все объясняющие переменные, а затем на последующих шагах рассматривать наборы объясняющих переменных, содержащие на одну переменную меньше, чем на предыдущем шаге.

4. Для работы с методом главных компонент необходимо загрузить пакет *dplyr*.
Для дальнейшей работы отберем только объясняющие переменные:

```
m<-select(ln_df, -mpg)
```

Применение метода главных компонент осуществляется функцией *prcomp*. Поскольку переменные разнородны, то перед применением метода главных компонент требуется осуществить их масштабирование. Это указывается переменной scale (scale = True) в аргументах функции *prcomp*. Применим теперь метод главных компонент и сохраним результаты в переменной m_pca:

```
m_pca<-prcomp(m, scale=T)
```

Применяя функцию *summary(m_pca)*, получим обобщающую информацию по результатам применения метода главных компонент: стандартные отклонения всех главных компонент, доли дисперсии объясняемой переменной, приходящейся на каждую из главных компонент и накопленные доли той же дисперсии. Для наглядного отображения этих дисперсий в виде гистограммы, можно воспользоваться командой *plot(m_pca).*

С помощью команды *m_pca$rotation* можно вывести матрицу весовых коэффициентов, с которыми каждая из объясняющих переменных входит в соответствующую главную компоненту.