Министерство науки и высшего образования РФ

ФГБОУ ВО «Ульяновский государственный университет»

Институт экономики и бизнеса

МЕТОДИЧЕСКИЕ УКАЗАНИЯ И ЗАДАНИЯ К ЛАБОРАТОРНЫМ РАБОТАМ ПО ДИСЦИПЛИНЕ «ЭКОНОМЕТРИЧЕСКОЕ МОДЕЛИРОВАНИЕ»

Ульяновск 2017

Методические указания и задания к лабораторным работам по дисциплине «Эконометрическое моделирование» / составитель: А.Е.Эткин.- Ульяновск: УлГУ, 2017.

Настоящие методические указания предназначены для студентов направления 38.03.05 «Бизнес-информатика». Указания необходимо использовать при выполнении лабораторных работ, предусмотренных учебным планом

Рекомендованы к использованию ученым советом
Института экономики и бизнеса УлГУ
Протокол № от « «_____ 2017г.

Лабораторная работа № 1.

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 № 2.

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 №1.

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 № 3.

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.