МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное АВТОНОМНОЕ образовательное учреждение высшего профессионального образования
«Национальный исследовательский ядерный университет «МИФИ»

**Димитровградский инженерно-технологический институт –**филиал федерального государственного автономного образовательного учреждения высшего
профессионального образования «Национальный исследовательский ядерный университет «МИФИ»
**(ДИТИ НИЯУ МИФИ)**

**Лабораторная работа №5
по курсу «Основы математической статистики и
планирование эксперимента»**Множественная линейная регрессия. Мультиколлинеарность

Составил: доцент кафедры
 высшей математики
 канд. экон. наук
 Кожухова В.Н.

Димитровград 2017

***Содержание***

[Цель работы 3](#_Toc494894659)

[Задание 3](#_Toc494894660)

[Теоретическая часть 4](#_Toc494894661)

[Модели множественной регрессии 4](#_Toc494894662)

[Идентификация моделей множественной линейной регрессии (МНК) 5](#_Toc494894663)

[Мультиколлинеарность. Корреляционная матрица 7](#_Toc494894664)

[Проверка качества уравнения множественной регрессии. Отбор факторов 9](#_Toc494894665)

[Пример выполнения работы 11](#_Toc494894666)

[Задание 11](#_Toc494894667)

[Ход работы 12](#_Toc494894668)

[Выводы 14](#_Toc494894669)

[Оформление отчета 15](#_Toc494894670)

[Контрольные вопросы 15](#_Toc494894671)

## Цель работы

Научиться осуществлять отбор факторов и оценивать значения параметров множественной линейной регрессии; выявлять мультиколлинеарность факторов; оценивать качество модели множественной регрессии.

## Задание

Имеются выборки равного объема для показателей *Y* и *X*1*, X*2*, X*3*, X*4. Предполагается наличие линейной зависимости уровней *Y* от *X*1*, X*2*, X*3*, X*4. Необходимо выполнить следующее.

1. Корреляционный анализ.

Построить корреляционную матрицу, проанализировать ее и отобрать неколлинеарные факторы.

1. Осуществить идентификацию модели со всеми включенными факторами, и модели с выбранными факторами.
2. Проверка качества полученной модели – расчет классического и скорректированного коэффициента детерминации
3. Расчет коэффициентов эластичности.

Рассчитать средние коэффициенты эластичности для каждого фактора, включенного в модель, пояснить их смысл.

## Теоретическая часть

Модели множественной регрессии

Уравнение множественной регрессии описывает зависимость результативного признака  от нескольких факторов .

Уравнение множественной линейной регрессии имеет вид:

.

Как и в случае парной регрессии, для построения модели необходимо решить задачи ее спецификации, идентификации и верификации.

В общем случае спецификация модели множественной регрессии включает:

* отбор факторов модели;
* выбор функционального вида модели.

Модель множественной регрессии может быть и нелинейной, как по переменным, так и по параметрам, например:

* логарифмическая 
* степенная ,
* экспоненциальная ,
* гиперболические или ,
* смешанные  и др.

В данной лабораторной работе рассматривается только линейная модель.

Что касается нелинейных, то для них справедливо все то же, что и в случае парной регрессии – нелинейные по параметрам модели необходимо линеаризовать, учитывая вхождение стохастической компоненты.

Широко используется модель множественной регрессии в *стандартизованном* масштабе:

,

,  – нормированные и центрированные величины: , , , .

Уравнение регрессии в стандартизованном масштабе не содержит свободного члена, а переменные ,  выражаются в долях от своих СКО.

Параметры (или их оценки) стандартизованного уравнения регрессии связаны с параметрами в *естественном* масштабе следующими соотношениями:

 

Значения параметров  сравнимы между собой, а параметров  – нет. Но  имеют качественную интерпретацию: каждый коэффициент линейной регрессии показывает, на сколько изменится зависимая переменная при увеличении соответствующей независимой переменной на 1 (и неизменности других переменных).

Идентификация моделей множественной линейной регрессии (МНК)

Идентификация модели множественной линейной регрессии осуществляется с помощью МНК.

.

Идентификацию можно выполнить несколькими способами, через уравнение регрессии в естественном масштабе или в стандартизованном масштабе.

I способ.

Решение СЛАУ, являющейся реализацией МНК:



II способ.

В матричном виде уравнение множественной регрессии имеет вид:

   

 .

Тогда МНК записывается в виде:



Его решение:





.

Замечание: под МНК нередко понимают именно эту формулу.

III способ.

Уравнению регрессии в стандартизованном масштабе соответствует следующая система:



Каждая сумма в данной системе преобразуется к соответствующему коэффициенту корреляции, например:

.

Таким образом:



Например, для 

.

При любом способе решения, для множественной линейной регрессии справедлива ***теорема Гаусса-Маркова***. Однако к пяти условиям Гаусса-Маркова, накладываемым на уравнение парной регрессии, добавляется еще одно:

6. Отсутствие *мультиколлинеарности*, т.е. линейной зависимости между объясняющими переменными:

.

При нарушении этого условия оценки параметров перестают быть эффективными, а решение системы уравнений в целом становится неустойчивым.

Можно сказать, что наличие линейной зависимости между факторами не позволяет «разделить» их влияние на эндогенную переменную и правильно рассчитать стоящие при них коэффициенты.

При *совершенной (строгой*) мультиколлинеарности между факторами существует явная функциональная зависимость:

.

На практике наиболее распространена *несовершенная* мультиколлинеарность, т.е. корреляционная зависимость между факторами:

.

Для устранения мультиколлинеарности используются следующие приемы:

1. Исключение факторов из модели. Если между двумя факторами существует мультиколлинеарность, один из них, менее информативный, следует исключить.
2. Замена переменных – переход от исходных данных к их разностям, темпам роста и т.п.
3. Изменение формы модели – переход от линейной зависимости к нелинейной. Это возможно, только если для нелинейной модели сохраняется зависимость между факторами и регрессором.
4. Получение новой выборки. На практике это не всегда возможно, но на другой выборке показатели могут оказаться некоррелированными.

Мультиколлинеарность. Корреляционная матрица

Проверка наличия мультиколлинеарности осуществляется путем анализа матрицы *парных* коэффициентов корреляции:

.

При отсутствии мультиколлинеарности корреляционная матрица должна иметь вид:

.

При совершенной мультиколлинеарности:

.

Таким образом, чем ближе определитель матрицы *R* к 0, тем выше мультиколлинеарность. Если  близок к 1, то мультиколлинеарность отсутствует.

Более строго проверку можно выполнить с помощью критерия .

 – отсутствие мультиколлинеарности;

 – наличие мультиколлинеарности.

Рассчитывается критерий, имеющий распределение  с  степенями свободы:

.

Если , то гипотеза  отклоняется, и в модели присутствуют коррелирующие факторы.

Факторы, оказывающие наибольшее влияние друг на друга, и наименьшее на результативный признак, необходимо исключить из модели. Можно по одному исключать «наихудшие» факторы до тех пор, пока мультиколлинеарность не исчезнет.

Недостаток парных коэффициентов корреляции заключается в том, что они не учитывают косвенное влияние факторов друг на друга.

Необходимо рассчитывать *частные* коэффициенты корреляции, которые очищены от влияния других факторов.

Частный коэффициент корреляции, очищенный от влияния *одного* фактора *xk* рассчитывается по формуле:



Частные коэффициенты корреляции, очищенные от влияния *всех* факторов рассчитываются через обратную матрицу :

.

Таким образом, получают матрицу частных коэффициентов корреляции:

.

Можно вычислить и частные коэффициенты корреляции между результативным признаком и факторами, например:



Частные коэффициенты корреляции, очищенные ото всех факторов, можно получить также можно получить через обратную матрицу *C*, но тогда в корреляционную матрицу нужно добавить строку и столбец для *Y*:

.

Частные коэффициенты корреляции позволяют судить о взаимосвязи между двумя переменными при фиксированных значениях других переменных.

Проверка качества уравнения множественной регрессии. Отбор факторов

Качество уравнения регрессии может быть проверено с помощью тех же показателей, что и для парной регрессии – MAE, MAPE-оценки, коэффициента детерминации и др.

В данной работе используются следующие критерии.

Коэффициент множественной детерминации:

.

Скорректированный коэффициент детерминации:

.

**F-критерий Фишера**:



Если , то нулевую гипотезу  следует отклонить, и принять модель и  статистически значимыми и надежными.

**t-статистика**:

,



Если , то соответствующий параметр можно считать статистически значимыми и надежными.

**Отбор факторов**, включаемых в уравнение регрессии, можно выполнить двумя путями:

1. последовательным *включением* факторов в модель (**пошаговый отбор факторов**) – сначала в модель включается один наиболее значимый фактор, затем второй и т.д., пока добавление новых факторов в модель повышает ее качество (скорректированный *R*2).
2. последовательным *исключением* факторов – сначала в модель включаются все факторы, затем наименее информативные по одному исключаются из модели, пока не начнет уменьшаться ее качество.

Какой путь выбрать – зависит от конкретной задачи. Если большинство из рассматриваемых факторов достаточно сильно коррелированны с регрессором, то проще идти методом исключения. Если факторов с достаточно сильной корреляцией немного, то удобнее применить пошаговый отбор.

В целом, при выборе факторов, которые могут быть потенциально включены в модель, необходимо, чтобы они обладали двумя *свойствами*: 1) были количественно измеримыи 2) не были коррелированны между собой.

## Пример выполнения работы

### Задание

Исследуется взаимосвязь показателей качества жизни населения по выборке для 25 регионов:

|  |  |
| --- | --- |
| Y | Средняя ожидаемая продолжительность жизни при рождении, лет |
| X1 | Уровень рождаемости, чел. на 1000 чел. населения |
| X2 | Доля населения с денежными доходами ниже величины прожиточного минимума, % от всего населения |
| X3 | Среднедушевые доходы населения, у.е. |
| X4 | Объем социальных выплат, млрд. у.е. |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| k | Y | X1 | X2 | X3 | X4 |
| 1 | 68,1 | 10,2 | 11,2 | 14,04 | 6,09 |
| 2 | 68,2 | 10,5 | 14,0 | 16,27 | 6,79 |
| 3 | 69,0 | 11,7 | 11,9 | 23,41 | 4,50 |
| 4 | 68,2 | 11,3 | 12,0 | 16,41 | 4,71 |
| 5 | 66,6 | 8,8 | 14,3 | 11,25 | 5,72 |
| 6 | 68,6 | 11,9 | 11,0 | 21,22 | 4,69 |
| 7 | 68,3 | 11,4 | 11,3 | 14,72 | 6,11 |
| 8 | 67,3 | 9,0 | 14,3 | 11,31 | 6,65 |
| 9 | 68,6 | 11,4 | 12,6 | 23,04 | 5,18 |
| 10 | 68,4 | 12,0 | 12,5 | 21,67 | 5,41 |
| 11 | 69,1 | 11,1 | 10,5 | 20,80 | 5,83 |
| 12 | 69,1 | 12,3 | 11,2 | 21,55 | 4,85 |
| 13 | 68,8 | 12,0 | 12,5 | 18,08 | 5,57 |
| 14 | 68,7 | 12,5 | 13,0 | 19,81 | 5,58 |
| 15 | 68,6 | 11,2 | 15,1 | 16,16 | 6,52 |
| 16 | 68,6 | 12,5 | 12,8 | 18,87 | 5,70 |
| 17 | 69,0 | 12,2 | 12,2 | 22,43 | 5,72 |
| 18 | 68,5 | 10,5 | 13,9 | 17,06 | 6,84 |
| 19 | 67,9 | 10,9 | 12,9 | 20,53 | 5,43 |
| 20 | 69,7 | 13,1 | 11,8 | 23,49 | 6,02 |
| 21 | 68,5 | 10,4 | 11,6 | 21,98 | 5,11 |
| 22 | 68,6 | 11,9 | 13,1 | 19,48 | 5,34 |
| 23 | 68,3 | 12,5 | 12,1 | 21,30 | 4,95 |
| 24 | 67,0 | 8,1 | 15,2 | 11,22 | 7,43 |
| 25 | 68,0 | 10,1 | 12,3 | 20,33 | 6,06 |

### Ход работы

#### Корреляционный анализ

Необходимо исследовать корреляционные зависимости между переменными. Составим матрицу парных коэффициентов корреляции:

.

Рассчитаем также парные коэффициенты корреляции между *Y* и *Xi*.

|  |  |
| --- | --- |
|  | Корреляционную матрицу можно получить так.Первый – через пакет «Анализ данных» → «Корреляция». |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Y** | **X1** | **X2** | **X3** | **X4** |
| **Y** | 1 | 0,8436 | -0,563 | 0,7874 | -0,382 |
| **X1** | 0,8436 | 1 | -0,522 | 0,7532 | -0,57 |
| **X2** | -0,563 | -0,522 | 1 | -0,579 | 0,635 |
| **X3** | 0,7874 | 0,7532 | -0,579 | 1 | -0,639 |
| **X4** | -0,382 | -0,57 | 0,635 | -0,639 | 1 |

Видим, что между всеми факторами существует корреляционная связь различной силы (от 0,5 до 0,75). Проверим наличие мультиколлинеарности с помощью определителя .

.

Определитель матрицы достаточно близок к нулю.

Рассчитаем частные коэффициенты корреляции с помощью обратной матрицы :

|  |  |  |
| --- | --- | --- |
| C |  | R0 |
|   | X1 | X2 | X3 | X4 |  | i\j  | 1 | 2 | 3 | 4 |
| X1 | 2,407 | 0,205 | -1,522 | 0,269 |  | 1 | -1,000 | -0,097 | **0,585** | -0,120 |
| X2 | 0,205 | 1,849 | 0,407 | -0,797 |  | 2 | -0,097 | -1,000 | -0,179 | 0,406 |
| X3 | -1,522 | 0,407 | 2,810 | 0,669 |  | 3 | **0,585** | -0,179 | -1,000 | -0,276 |
| X4 | 0,269 | -0,797 | 0,669 | 2,087 |  | 4 | -0,120 | 0,406 | -0,276 | -1,000 |

|  |  |  |  |
| --- | --- | --- | --- |
| Cy |  | i | ryxi. |
| 6,614 | -4,059 | 1,605 | -2,934 | -2,681 |  |   |   |
| -4,059 | 4,898 | -0,780 | 0,279 | 1,915 |  | 1 | 0,713 |
| 1,605 | -0,780 | 2,239 | -0,305 | -1,448 |  | 2 | -0,417 |
| -2,934 | 0,279 | -0,305 | 4,111 | 1,858 |  | 3 | 0,563 |
| -2,681 | 1,915 | -1,448 | 1,858 | 3,174 |  | 4 | 0,585 |

Наибольшее влияние на результативный признак Y оказывает фактор X1, наименьшее X2. Мультиколлинеарность в наибольшей степени вызвана корреляцией между X1 и X3. Следовательно, один из них необходимо исключить из модели. В данном случае лучше исключить X3, поскольку он оказывает меньшее влияние на Y. Высокое значение парного коэффициента корреляции между Y и X3 обусловлено именно мультиколлинеарностью.

#### Отбор факторов и идентификация модели

Выполним идентификацию модели с включением переменных X1, X2 и X4.

.

Параметры вычислим через матричное представление исходных данных.

.

|  |
| --- |
| XTX |
| 25 | 279,392 | 315,425 | 142,831 |
| 279,392 | 3160,13 | 3505,13 | 1583,48 |
| 315,425 | 3505,13 | 4018,38 | 1816,48 |
| 142,831 | 1583,48 | 1816,48 | 829,291 |
| Z = (XTX)-1 |
| 18,4455 | -0,77039 | -0,48326 | -0,64738 |
| -0,77039 | 0,04192 | 0,01115 | 0,02822 |
| -0,48326 | 0,01115 | 0,04631 | -0,03948 |
| -0,64738 | 0,02822 | -0,03948 | 0,14531 |
| XTY |  | B = Z(XTY) |  |
| 1709,8 |  | 63,8283 |  |
| 19125,1 |  | 0,45287 |  |
| 21561,2 |  | -0,16085 |  |
| 9763,96 |  | 0,26815 |  |

|  |  |  |
| --- | --- | --- |
|  | XT | =ТРАНСП(матрица\_X) |
|  | XTX | =МУМНОЖ( ТРАНСП(матрица\_X); матрица\_X) |
|  | (XTX)-1 | =МОБР(матрица\_XTX) |
|  | Все формулы необходимо использовать как формулы массивов (выделить диапазон, F2, Ctrl+Shift+Enter). |

Таким образом, модель имеет вид:

.

Коэффициент детерминации:

.

.

Окончательно модель имеет вид

.

Полученная модель описывает 77% исходных данных, 23% приходятся на случайные отклонения.

Самостоятельно реализуйте идентификацию модели со всеми включенными факторами.

#### Проверка качества полученной модели

Проверим общее качество модели с помощью F-теста:



, следовательно,  и полученная модель статистически значимы и надежны.

### Выводы

Таким образом, получена зависимость средней ожидаемой продолжительности жизни при рождении от уровня рождаемости, доли малоимущего населения (с доходом менее прожиточного минимума) и объема социальных выплат:

.

Фактор среднедушевого дохода был исключен из модели из-за сильной коррелированности с другими факторами.

## Оформление отчета

Отчет по лабораторной работе оформляется в текстовом процессоре MS Word или аналогичном и сдается в печатном виде. Рекомендуется использовать шрифт Times New Roman или Cambria, 14пт., абзацный отступ 1,25см, межстрочный интервал – 1 или 1,5, поля: верхнее 1,5-2см, нижнее 2см, правое 1-1,5см, левое 2,5-3см. Страницы нумеруются снизу по центру, кроме титульного листа.

Отчет должен содержать:

1. Титульный лист.
2. Индивидуальное задание.
3. Основная часть.

Результаты выполнения каждого пункта задания с краткими комментариями и указанием расчетных формул. При необходимости приводить основные промежуточные расчеты.

1. Вывод.

Привести и сопоставить основные численные результаты работы.

## Контрольные вопросы

1. Каков смысл коэффициентов множественной регрессии в естественном масштабе.
2. В каких величинах выражаются переменные в модели множественной регрессии в стандартизованном масштабе?
3. По результатам идентификации получены значения , . Какой из факторов оказывает большее влияние на результативный признак?
4. Что показывают средние коэффициенты эластичности?
5. Запишите решение МНК в матричном виде.
6. Запишите систему для решения МНК через коэффициенты корреляции.
7. Какое условие Гаусса-Маркова добавляется к условиям для парной линейной регрессии? Каковы последствия его нарушения?
8. Что такое корреляционная матрица?
9. В каком диапазоне изменяется определитель корреляционной матрицы ?
10. Что можно сказать о модели множественной регрессии, для которой ? ?
11. Чем частные коэффициенты корреляции отличаются от парных?
12. Как рассчитывается скорректированный коэффициент детерминации? Как он соотносится с обычным ?
13. Какими основными свойствами должны обладать факторы модели множественной регрессии?