МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное бюджетное образовательное учреждение
высшего образования
«Тольяттинский государственный университет»

Институт математики, физики и информационных технологий
(наименование института полностью)

Кафедра «Прикладная математика и информатика»
(наименование кафедры)

01.03.02 Прикладная математика и информатика
(код и наименование направления подготовки, специальности)

Системное программирование и компьютерные технологии
(направленность (профиль)/специализация)

БАКАЛАВРСКАЯ РАБОТА

на тему Алгоритм обобщенной множественной линейной регрессии и его реализация

Студент	А.Б. Анорова	
	(И.О. Фамилия)	(личная подпись)
Руководитель	Г.А. Тырыгина	
	(И.О. Фамилия)	(личная подпись)
Консультанты	Е.В. Косс	
	(И.О. Фамилия)	(личная подпись)

Допустить к защите

Заведующий кафедрой к.т.н., доцент, А.В. Очеповский _____
(степень, звание, И.О. Фамилия) (личная подпись)

«_____»_____20_____г.

Тольятти 2018

АННОТАЦИЯ

Темой выпускной квалификационной работы является: «Алгоритм обобщенной множественной линейной регрессии и его реализация».

Объект исследования: математическая модель множественной регрессии.

Предмет исследования: математическая модель множественной регрессии в условиях нарушения предпосылок классической линейной модели множественной регрессии (ЛММР).

Целью работы является: разработка и реализация алгоритма обобщенной модели множественной регрессии.

Для достижения цели работы необходимо решить следующие задачи:

1. Проанализировать алгоритм классической линейной модели множественной регрессии.
2. Проанализировать регрессионную модель в условиях нарушения классических предпосылок.
3. Осуществить анализ реализаций линейной множественной регрессионной модели.

Дипломная работа состоит из введения, трех глав и заключения.

В первой главе данной работы описана и проанализирована классическая линейная модель множественной регрессии.

Во второй главе представлены описание и анализ обобщенных линейных моделей множественной регрессии.

В третьей главе представлен сравнительный анализ построения двух описанных выше моделей на примере одних и тех же данных.

В заключении представлены выводы и результаты проделанной работы.

Пояснительная записка содержит 40 страницы, содержит введение, три главы, заключение и список литературы, состоящий из 27 литературных источников, 5 рисунка и 1 таблицу.

Результатом работы является сравнительный анализ классической и обобщенной моделей линейной множественной регрессии.

ABSTRACT

The title of the graduation work is «Generalized multiple linear regression algorithm and its implementation».

The object of the graduation work is a mathematical model of multiple regression.

The subject of the graduation work is a mathematical model of multiple regression under conditions of violation of the classical linear model of multiple regression (LMMR).

The aim of the work is to give some information about the development and implementation of the generalized multiple regression model.

To achieve the goal, we will form the following tasks:

1. To analyze the algorithm of the classical LMMR.
2. To analyze the regression model under conditions of violation of classical assumptions.
3. To analyze the implementation of the LMMR.

The graduation work consists of an introduction, three chapters and a conclusion.

The first Chapter of this paper describes and analyzes classic LMMR.

The second Chapter describes and analyzes generalized LMMR.

The third Chapter presents the comparative analysis of the construction of the two models described above as exemplified by the same data.

Finally, the conclusions and results of the work are presented.

The graduation work consists of an explanatory note on 40 pages, introduction, three chapters, a conclusion 5 figures and a list of 27 references.

The result of this work is the comparative analysis of the classical and generalized models of linear multiple regression.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Регрессионный анализ – это статистический инструмент, который применяется с целью определения взаимосвязи между двумя или более количественными переменными и прогнозирования значений зависимых переменных. Раньше, до появления компьютера применение этого метода было затруднительно, особенно при больших объемах данных.

На практике условия регрессионной модели выполняются крайне редко. Поэтому актуально построение математической модели при более реалистичных предположениях с несмещенными и состоятельными оценками. В данной работе описан обобщенный метод наименьших квадратов, который помогает решать задачи данного типа.

Целью данной работы является разработать и реализовать алгоритм обобщенной модели множественной регрессии.

Объектом является математическая модель множественной регрессии.

Предметом является математическая модель множественной регрессии в условиях нарушения предпосылок классической линейной модели множественной регрессии.

Для реализации были сформулированы следующие задачи:

1. Проанализировать алгоритм классической линейной модели множественной регрессии.
2. Проанализировать регрессионную модель в условиях нарушения классических предпосылок.
3. Осуществить анализ реализаций линейной множественной регрессионной модели.

Структура. Дипломная работа состоит из введения, двух глав, заключения и приложения.

В первой главе данной работы описана и проанализирована классическая линейная модель множественной регрессии.

Во второй главе представлены описание и анализ обобщенных линейных моделей множественной регрессии.

В третьей главе представлен сравнительный анализ построения двух описанных выше моделей на примере одних и тех же данных.

В заключении представлены выводы и результаты проделанной работы.

1 КЛАССИЧЕСКАЯ ЛИНЕЙНАЯ МОДЕЛЬ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

1.1 Алгоритм классической линейной модели множественной регрессии

Модель множественной регрессии позволяет находить и изучать зависимости переменной Y от нескольких объясняющих переменных $X_1, X_2, \ldots, X_n$ [1].

Данные модели являются наиболее важными и широко используемыми статистическими методами. Чаще всего они применяются в моделировании и прогнозировании: исследование и установление связи между набором переменных [4].

Множественный регрессионный анализ используется с целью нахождения статистически значимой связи между наборами переменных. Он используется для поиска тенденций в этих наборах данных. Множественный регрессионный анализ аналогичен парной линейной регрессии. Единственным отличием между парной линейной регрессией и множественной регрессией является количество предикторов (переменных), используемых в регрессии [2].

Парный регрессионный анализ использует одну переменную *X* для каждой зависимой переменной *Y*. Например: (X_1, Y_1). Множественная регрессия использует несколько переменных для каждой независимой переменной [3].

Пусть *i*-е наблюдение объясняющих переменных $\chi_{i1}, \chi_{i2}, \ldots, \chi_{ip}$, а γ_i — зависимой переменной. Вследствие этого модель множественной линейной регрессии представляется в виде:

$$\gamma_i = \beta_0 + \beta_1 \chi_{i1} + \beta_2 \chi_{i2} + \cdots + \beta_p \chi_{ip} + \varepsilon_i, \quad (1.1)$$

где $i = 1,2,\ldots,n$; ε_i – случайная составляющая модели [18].

Переменные γ и ε являются реализациями случайных величин Υ и X. Единственным источником неопределенности (случайности) в уравнении (1.1) становится случайная составляющая модели ε. Переменная $\chi_{\mathbf{1}} = \chi_{i\mathbf{1}} = 1$ (для всех $= 1,2,\ldots,n$) называется вспомогательной переменной для свободного члена. Эту переменную вводят для удобства записи множественной линейной регрессии. Коэффициент регрессии β_p является свободным членом и называется параметром сдвига.

Данная модель множественной линейной регрессии будет являться обобщением модели парной линейной регрессии на многомерный случай [21].

В том случае, когда в модель регрессии вводятся новые переменные – модель усложняется, а вместе с ней и получаемые формулы. Таким образом, рациональнее использовать матричные обозначения. Теоретические концепции анализа и необходимые процедуры расчета можно облегчить матричным описанием регрессии [10].

Пусть $\Upsilon = (\gamma_{\mathbf{1}}\gamma_2 \ldots \gamma_n)'$— матрица-столбец, или вектор, значений зависимой переменной размера $n^{\mathbf{1}}$;

$$\mathrm{X} = \begin{matrix} 1 & \chi_{\mathbf{11}} & \chi_{\mathbf{12}} & \cdots & \chi_{\mathbf{1}p} \\ 1 & \chi_{21} & \chi_{22} & \cdots & \chi_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & \chi_{n\mathbf{1}} & \chi_{n2} & \cdots & \chi_{np} \end{matrix}$$

— матрица значений пояснительных переменных или матрица плана размера $n \times\ p + 1$;

$\beta = \beta_0\beta_{\mathbf{1}} \ldots \beta_p\ '$— матрица-столбец или вектор, параметры измерения $p + 1$;

$\varepsilon = (\varepsilon_{\mathbf{1}}\varepsilon_2 \ldots \varepsilon_n)'$— матрица-столбец или вектор, возмущения размера n.

Тогда в матричной форме модель (1.1) принимает вид:

$$\Upsilon = \mathrm{X}\beta + \varepsilon. \qquad (1.2)$$

Оценка этой модели – уравнение

$$Y = Xb + e, \qquad (1.2')$$

где $b = \begin{pmatrix} b_0 b_1 \dots b_p \end{pmatrix}', e = \begin{pmatrix} e_1 e_2 \dots e_n \end{pmatrix}'$.

Об изменениях величины зависимой переменной Y, при увеличении объясняющей переменной X, можно судить по коэффициенту регрессии β [20].

Для того чтобы оценить параметры модели, изучить свойства и определить качество построенной модели необходимо определить предпосылки регрессионной модели:

1. Независимые переменные не случайны и измеряются без ошибок (матрица X – детерминированная).

2. Математическое ожидание возмущения ε_i или зависимой переменной y_i равно нулю:

$$M(\varepsilon_i) = 0$$
$$M(y_i) = \beta_0 + \beta_1 x_i).$$

3. Дисперсия возмущения ε_i или зависимой переменной y_i одинакова для любого i:

$$D(\varepsilon_i) = \sigma^2.$$

Данное свойство означает, что возмущения или зависимая переменная гомоскедастичны.

4. Возмущения ε_i и ε_j (или переменные y_i и y_j) некоррелированы (независимы) для различных наблюдений:

$$M(\varepsilon_i \varepsilon_j) = 0 \ \ i \neq j \ .$$

5. Случайная составляющая и объясняющие переменные некоррелированы. Для классической регрессионной модели данная предпосылка будет выполняться всегда, т.к. первая предпосылка говорит о детерминированности объясняющих переменных.

6. Коэффициенты регрессии являются постоянными величинами.

7. Регрессоры (объясняющие переменные) не коллинеарны. Данная предпосылка говорит о существования и единственности решения задачи оценивания параметров модели по МНК.

8. Возмущение ε_i (или зависимая переменная y_i) является нормально распределенной случайной величиной.

9. Количество наблюдений больше количества оцениваемых параметров.

Итак, в том случае, когда зависимая переменная γ_i, возмущения ε_i и объясняющие переменные $\chi_{i1}, \chi_{i2}, \ldots, \chi_{ip}$.. удовлетворяют вышеприведенным предпосылкам регрессионного анализа линейная модель множественной регрессии (ЛММР) будет называться классической нормальной моделью [11].

1.2 Базовый анализ классической модели множественной регрессии

Метод наименьших квадратов применяется для оценки вектора неизвестных параметров β [17]. Запишем условие минимизации для остаточной суммы квадратов в виде:

$$S = \sum_{i=1}^{n} (\gamma_{\chi i} - \gamma_i)^2 = \sum_{i=1}^{n} e_i^2 = e'e = (\Upsilon - \mathrm{X}b)'(\Upsilon - \mathrm{X}b) \to min. \qquad (1.3)$$

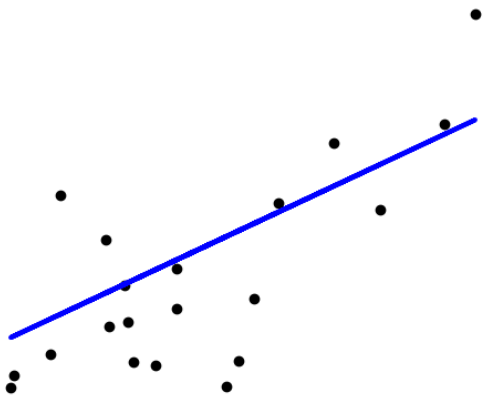

Рисунок 1 – Минимизация остаточной суммы

После раскрытия скобок получаем:

$$S = \mathrm{Y}'\mathrm{Y} - b'\mathrm{X}'\mathrm{Y} - \mathrm{Y}'\mathrm{X}b + b'\mathrm{X}'\mathrm{X}b.$$

Учитывая, что $\mathrm{Y}'\mathrm{X}b = \ \mathrm{Y}'\mathrm{X}b\ ' = b'\mathrm{X}'\mathrm{Y}$, т.к. величины в правой и левой части – скаляры, условие минимизации (1.3) примет вид:

$$S = \mathrm{Y}'\mathrm{Y} - 2b'\mathrm{X}'\mathrm{Y} + b'\mathrm{X}'\mathrm{X}b \to min. \qquad (1.4)$$

Функция $S\ b_0, b_1, \ldots, b_p$ представляет собой квадратичную форму относительно вектора оценок b. Найдем его экстремум, приравнивая к нулю частные производные функции $S\ b_0, b_1, \ldots, b_p$ относительно этих переменных. Запишем вектор частных производных в матричной форме:

$$\frac{\partial S}{\partial b} = \ \frac{\partial S}{\partial b_0}\frac{\partial S}{b_1}\cdots\frac{\partial S}{b_p}\ .$$

Для определения вектора b получим систему нормальных уравнений в матричной форме:

$$\mathrm{X}'\mathrm{X}b = \mathrm{X}'\mathrm{Y}. \qquad (1.5)$$

Представим данную систему в развернутом виде:

$$\mathrm{X}'\mathrm{X} = \begin{matrix} 1 & 1 & 1 & \ldots & 1 \\ \chi_{11} & \chi_{21} & \chi_{32} & \cdots & \chi_{n1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \chi_{1p} & \chi_{2p} & \chi_{3p} & \cdots & \chi_{np} \end{matrix} \begin{matrix} 1 & \chi_{11} & \chi_{12} & \cdots & \chi_{1p} \\ 1 & \chi_{21} & \chi_{22} & \cdots & \chi_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & \chi_{n1} & \chi_{n2} & \cdots & \chi_{np} \end{matrix} = \qquad (1.6)$$

$$= \begin{matrix} n & \chi_{i1} & \cdots & \chi_{ip} \\ \chi_{i1} & {\chi_{i1}}^2 & \cdots & \chi_{i1}\chi_{ip} \\ \cdots & \cdots & \cdots & \cdots \\ \chi_{ip} & \chi_{i1}\chi_{ip} & \cdots & {\chi_{ip}}^2 \end{matrix}$$

Матрица $\mathrm{X}'\mathrm{Y}$ - вектор произведений n наблюдений объяснительных и зависимых переменных:

$$X'Y = \begin{matrix} 1 & 1 & 1 & \dots & 1 \\ x_{11} & x_{21} & x_{32} & \dots & x_{n1} \\ \dots & \dots & \dots & \dots & \dots \\ x_{1p} & x_{2p} & x_{3p} & \dots & x_{np} \end{matrix} \begin{matrix} y_1 \\ y_2 \\ \dots \\ y_n \end{matrix} = \begin{matrix} y_i \\ y_i x_{i1} \\ \dots \\ y_i x_{ip} \end{matrix} \qquad (1.7)$$

Умножая матрицы и векторы в выражениях (1.6) и (1.7) в частном случае для одной объясняющей переменной *(p = 1)*, получим систему нормальных уравнений:

$$\begin{cases} b_0 n + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i ; \\ b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i . \end{cases} \qquad (1.8)$$

Матричное уравнение (1.5) принимает вид:

$$\begin{matrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{matrix} \begin{matrix} b_0 \\ b_1 \end{matrix} = \begin{matrix} \sum y_i \\ \sum y_i x_i \end{matrix} ,$$

из которого следует система нормальных уравнений (1.8).

Из предпосылки 5 множественного регрессионного анализа следует, что определитель матрицы $X'X$ равен нулю. Из этого следует, что ранг матрицы $X'X$ равен его порядку, т. е. $r\ X'X\ = p + 1$. Известно, что $r\ X'X\ = r\ X\ $, следовательно, $r\ X\ = p + 1$, т. е. ранг матрицы плана X равен числу его столбцов. Принимая во внимание приведенные выше объяснения, запишем предпосылку 5 регрессионного анализа следующим образом:

5. Для векторов значений объясняющих переменных должно выполняться свойство нелинейности, т. е. ранг матрицы X является максимальным ($r\ X\ = p + 1$).

Для того чтобы получить достоверные статистические выводы количество доступных наблюдений (значений) каждой из объясняющих и зависимых переменных должно превышать ранг матрицы X, т. е. $n > r\ X$ или $n > p + 1$.

N-мерным аналогом среднего отклонения от математического ожидания одной переменной является ковариационная матрица вектора возмущений $_{\varepsilon}$.

Как описано выше в п. 1.1, при выполнении предпосылок 1-9 модель называется классической нормальной моделью (КНЛММР). Если же предпосылка 8 о нормальном законе распределения вектора возмущений не выполняется, то модель носит название просто классической модели (КЛММР).

Решением уравнения (1.5) является вектор

$$b = \ X'X\ ^{-1}X'Y, \tag{1.9}$$

Теорема Гаусса-Маркова.

Если регрессионная модель удовлетворяет предпосылкам 1-4, то оценки b модели имеют наименьшую дисперсию в классе всех линейных несмещенных оценок.

Оценка метода наименьших квадратов $b = (X'X)^{-1}X'Y$ окажется наиболее эффективной в случае выполнения предпосылок множественного регрессионного анализа.

Для того чтобы правильно оценить влияние объясняющих переменных на зависимую, необходимо чтобы они были приведены к единым единицам измерения [7]. Для этого используются стандартизованные коэффициенты регрессии b'_j и коэффициенты эластичности $E_j\ \ j = 1,2,\ldots,p$:

$$b'_j = b_j \frac{s_{x_j}}{s_y}; \tag{1.11}$$

$$E_j = b_j \frac{\bar{x}_j}{\bar{y}}. \tag{1.12}$$

Точность уравнения множественной регрессии определяется по изменениям оценок параметров. В множественном регрессионном анализе используется аналог дисперсии одной переменной — ковариационная матрицу вектора оценки b:

$$\underset{b}{=} \begin{matrix} \sigma_{00} & \sigma_{01} & \cdots & \sigma_{0p} \\ \sigma_{10} & \sigma_{11} & \cdots & \sigma_{1p} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{p0} & \sigma_{p1} & \cdots & \sigma_{pp} \end{matrix}$$

где элементы σ_{ij} — ковариация оценок параметров β_i и β_j.

Таким образом:

$$\sigma_{ij} = M \quad b_i - M\ b_i \quad\quad b_j - M\ b_j \quad . \tag{1.14}$$

Ковариация – величина, показывающая совместное изменение двух величин. Это похоже на дисперсию, но в то время как дисперсия показывает изменение одной величины, ковариация показывает как изменяются переменные вместе [19].

Оценки b_j, которые были получены применением метода наименьших квадратов, будут несмещенными оценками параметров b_j, т. е. $M\ b_j\ = \beta_j$, выражение (1.13) принимает вид:

$$\underset{b}{=} M \quad b - \beta\ \ (b - \beta)' \tag{1.15}$$

Учитывая, (1.12), преобразуем это выражение:

$$\begin{aligned} {}_b = M \quad X'X\ ^{-1}X'\varepsilon \quad X'X\ ^{-1}X'\varepsilon\ ' = M \quad X'X\ ^{-1}X'\varepsilon\varepsilon'X\ X'X\ ^{-1} = \\ = \ X'X\ ^{-1}X'M\ \varepsilon\varepsilon'\ X\ X'X\ ^{-1}, \end{aligned} \tag{1.16}$$

элементы матрицы X не являются случайными величинами.

Матрица $M\ \varepsilon\varepsilon'$ - ковариационная матрица вектора возмущения:

$$_{\varepsilon} = M\ \varepsilon\varepsilon' = \begin{matrix} M(\varepsilon_1^2) & M(\varepsilon_1\varepsilon_2) & \cdots & M(\varepsilon_1\varepsilon_n) \\ M(\varepsilon_2\varepsilon_1) & M(\varepsilon_2^2) & \cdots & M(\varepsilon_2\varepsilon_n) \\ \cdots & \cdots & \cdots & \cdots \\ M(\varepsilon_n\varepsilon_1) & M(\varepsilon_n\varepsilon_2) & \cdots & M(\varepsilon_n^2) \end{matrix},$$

где элементы, лежащие на главной диагонали равны дисперсии σ^2:

$$M\ \varepsilon_i^2\ = M(\varepsilon_i - 0)^2 = D\ \varepsilon_i^2\ = \sigma^2,$$

а все остальные элементы равны нулю, т.к. возмущения некоррелированы между собой.

Следовательно, матрица

$$M\ \varepsilon\varepsilon'\ = \sigma^2 E_n,$$

где E_n — единичная матрица n-го порядка. Поэтому, в силу (1.16) ковариационная матрица оценок параметров:

$$_b = \sigma^2\ X'X\ ^{-1}. \qquad (1.17)$$

Дисперсию и ковариацию, а также вектор b оценок параметров можно определить с помощью обратной матрицы $X'X\ ^{-1}$.

2 ОБОБЩЕННАЯ ЛИНЕЙНАЯ МОДЕЛЬ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

2.1 Обобщенная линейная модель множественной регрессии при известной ковариационной матрице регрессионных остатков

В эмпирических социально-экономических исследованиях зачастую условия классической линейной модели нарушаются. Например, двоичные ответы (да/нет или 0/1) не имеют одинаковой дисперсии между классами. Кроме того, сумма членов в линейной модели обычно может иметь очень большие диапазоны, охватывающие отрицательные и положительные значения [8]. Для примера бинарного ответа мы хотели бы, чтобы ответ был вероятностью в диапазоне [0,1]. Обобщенные линейные модели учитывают реакции, нарушающие допущения линейной модели, с помощью двух механизмов: функции связи и функции дисперсии. Функция связи преобразует целевой диапазон в потенциально бесконечность до бесконечности, чтобы можно было поддерживать простую форму линейных моделей. Функция дисперсии выражает дисперсию как функцию предсказанного ответа, тем самым приспосабливая ответы с непостоянными дисперсиями (такими как двоичные ответы). В таких случаях классическая модель будет служить базой при построении так называемых обобщенных моделей [12]. Для оценивания параметров в этих моделях применяются методы, которые являются модификациями обычного метода наименьших квадратов. Метод наименьших квадратов (МНК, OLS, Ordinary Least Squares) — базовый метод регрессионного анализа, который помогает оценить неизвестные параметры модели, составленной по выборочным данным [5]. Перечислим возможные модификации МНК, построенные таким образом, чтобы соответствующие оценки оставались состоятельными и несмещенными.

1. Метод инструментальных переменных. Используется для построения несмещенных и состоятельных оценок для модели при

нарушении предпосылки о том что, независимые переменные не случайны и измеряются без ошибок (матрица X – детерминированная).

Нарушение данной предпосылки может привести к смещенности оценок метода наименьших квадратов. Это произойдет, если хотя бы один регрессор (независимая переменная) и случайная составляющая модели коррелированы (зависимы) между собой. Идея метода состоит в подборе новых независимых переменных (стохастических, случайных) таким образом, чтобы эти переменные были сильно коррелированы с регрессорами модели, но были не коррелированы с ее возмущениями.

2. Метод введения «фиктивных» переменных. Данный метод позволяет учесть все структурные изменения в случае, когда коэффициенты регрессии являются непостоянными величинами [23].

Значения коэффициентов $\beta_\rho, \mathrm{p} = 1,2,\ldots,k,$ одинаковы для всех элементов наблюдаемой пространственной или временной выборки. Однако, в одной и той же выборке могут содержаться данные о неоднородных объектах. В этом случае модель классической регрессии будет неадекватной, т.к. не будет соответствовать наблюдаемым данным.

3. При нарушении предпосылки о неколлинеарности регрессоров (объясняющих переменных), которая говорит о существовании и единственности решения задачи оценивания параметров модели по МНК, появляется проблема мультиколлинеарности регрессоров. Это приводит к таким проблемам как – увеличение дисперсий оценок коэффициентов, неидентифицируемость модели и неустойчивость оценок, увеличение доверительных интервалов.

Существует подход по устранению коллинеарности – изменение спецификации модели, посредством удаления из нее регрессора, который сильно коррелирует с другими. Но при отбрасывании существенной переменной может произойти нарушение предпосылки о правильной спецификации. Поэтому при удалении независимой переменной следует

учитывать экономический смысл переменных и степень их влияния на зависимую переменную.

4. При невыполнении предпосылок 3 и 4 модель регрессии принято называть обобщенной линейной моделью множественной регрессии (ОЛММР).

Рассмотрим подробно данную модель.

Допустим, что переменные и параметры удовлетворяют следующим условиям:

1. ε — случайный вектор возмущений; X — детерминированная матрица;

2. $M\ \varepsilon\ = 0_n$;

3. $\ _{\varepsilon} = M\ \varepsilon\varepsilon'\ = \Omega,\ \Omega$ — положительно определенная матрица;

4. $r\ X\ = p + 1 < n,\ p$ – число объясняющих переменных; n – число наблюдений.

Тогда ОЛММР можно записать в виде:

$$Y = X\beta + \varepsilon. \tag{2.1}$$

Обобщенные линейные модели (GLMRM) представляют собой широкий класс моделей, которые включают линейную регрессию, дисперсионный анализ, регрессию Пуассона, лог-нормальная модели и т. д. В таблице ниже представлены модели и их краткая характеристика:

Таблица 1 – обобщенные линейные модели

Модель	Распределение	Функция связи	Систематическая составляющая
Линейная регрессия	Нормальное	Линейная	Непрерывная
Дисперсионный анализ	Нормальное	Линейная	Категориальная
Ковариационный анализ	Нормальное	Линейная	Смешанная

Логистическая регрессия	Биномиальное	Логарифмическая	Смешанная
Логарифмическая регрессия	Пуассоновское	Логарифмическая	Категориальная
Регрессия Пуассона	Пуассоновское	Логарифмическая	Смешанная
Множественная регрессия	Полиномиальное	Обобщенная логарифмическая	Смешанная

Распределение – распределение зависимой переменной, например нормальное распределение для Y в линейной регрессии или биномиальное распределение в парной логистической регрессии. Также называют случайной составляющей модели.

Функция связи – описывает связь между случайными и систематическими компонентами. Показывает вид связи зависимой и объясняющих переменных, например, $\eta = g(E(Y_i)) = E(Y_i)$ – линейная, $\eta = logit\ (\pi)$- логистическая.

Систематическая составляющая – определяет объясняющие переменные $X_1, X_2, \ldots, X_n$ в модели, а точнее их линейную комбинацию при создании так называемого линейного предиктора.

Модели GLM генерируют следующую статистику коэффициентов:

- оценка линейных коэффициентов;
- стандартная погрешность оценки коэффициента;
- t - значение оценки коэффициента;
- вероятность t-значения;
- коэффициент вариации;
- стандартизированная оценка коэффициента;
- нижняя и верхняя доверительные границы коэффициента.

Однако оценки коэффициентов для обычных наименьших квадратов зависят от независимости переменных. Когда переменные коррелируют и столбцы матрицы имеют приближенную линейную зависимость, матрица

становится близкой к сингулярной и в результате оценка наименьших квадратов становится очень чувствительной к случайным ошибкам в наблюдаемом ответе, производя большую дисперсию. Такая ситуация мультиколлинеарности может возникнуть, например, при сборе данных без экспериментальной разработки.

Таким образом, видно, что ковариационные матрицы обобщенной и классической моделей будут различны: для классической – $_{\varepsilon} = \Omega = \sigma^2 E_n$, для обобщенной – $_{\varepsilon} = \Omega$.

Если применить к данной модели обычный метод наименьших квадратов (МНК) это приведет к следующим последствиям:

–Оценки коэффициентов модели не будут эффективными.

–МНК-оценка дисперсии случайной составляющей в обобщенной модели является смещенной;

– МНК-оценка ковариационной матрицы вектора оценок коэффициентов является смещенной оценкой истинной ковариационной матицы обобщенной модели;

Для обобщенной модели получим

$$_{b^*} = (X'X)^{-1}X'M\ \varepsilon\varepsilon'\ X\ X'X^{-1} = X'X^{-1}X'\Omega X\ X'X^{-1}, \qquad 2.2$$

А учитывая **(**1.17**)** для классической модели было:

$$_{b} = \sigma^2\ X'X^{-1}. \qquad 2.3$$

Математическое ожидание остаточной суммы квадратов равно $\sum_{i=1}^{n} e_i^2 = e'e$. Для обобщенной модели:

$$M\ e'e = tr\ E_n - X\ X'X^{-1}X'\ \Omega, \qquad 2.4$$

т. е. в соответствии с несмещенной оценкой s^2, которая определяется как

$$s^2 = \frac{e'e}{n-p-1} = \frac{\sum_{i=1}^{n} e_i^2}{n-p-1} \qquad 2.5$$

математическое ожидание можно выразить в виде:

$$M\ s^2 = M\ \frac{e'e}{n-p-1} = \frac{tr\ E_n - X\ X'X^{-1}X'\ \Omega}{n-p-1}, \qquad 2.6$$

где tr – след соответствующей матрицы.

В формуле (2.3) заменим оценку ковариационной матрицы $_b\ \sigma^2$ на s^2 и получим:

$$M\ _b = M\ s^2\ X'X^{-1} = \frac{tr\ E_n - X\ X'X^{-1}X'\ \Omega}{n-p-1}\ X'X^{-1} \qquad 2.7$$

Рассчитанное данным образом математическое ожидание в общем случае не будет совпадать с ковариационной матрицей, описанной выше, что говорит о смещенности полученной оценки. Оценка b, будет состоятельной, но не будет оптимальной. Чтобы получить наиболее эффективную оценку необходимо воспользоваться обобщенным методом наименьших квадратов.

Теорема Айткена.

Оценка вектора β

$$b^* = X'\Omega^{-1}X^{-1}X'\Omega^{-1}Y \qquad 2.8$$

обобщенной регрессионной модели в классе линейных несмещенных оценок имеет наименьшую ковариационную матрицу.

Наиболее эффективной оценкой, по теореме Гаусса-Маркова, является оценка (1.9), т. е.

$$b^* = (X_*'X_*)^{-1}X_*'Y_*. \qquad 2.14$$

Возвращаясь к исходным наблюдениям X и Y и учитывая (7.9), получим

$$b^* = \quad P^{-1}X\ '\ P^{-1}X\ \ ^{-1}\ P^{-1}X\ 'P^{-1}Y =$$
$$= \ X'\ P^{-1}\ 'P^{-1}X\ ^{-1}X'\ P^{-1}\ 'P^{-1}Y = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y,$$

т. е. выражение (2.7), что и требовалось доказать.

Итак, получается, что оценка, которая получена обобщенным методом наименьших квадратов b^*, равна оценке «обычного» метода b.

Метод максимального правдоподобия позволяет получить состоятельную оценку b^*, в случае, когда известна матрица Ω, и выполняется предпосылка о нормальном законе распределения ε.

Оценка $b^* = (X_*'X_*)^{-1}X_*'Y_*$ - точка минимума по b остаточной суммы квадратов, в соответствии с (1.3):

$$S = \ \sum_{i=1}^{n} e_{i*}^2 = e_*'e_* = (Y_* - X_*b)'(Y_* - X_*b).$$

Посмотрим на исходные наблюдения,

$$S = \ P^{-1}\ Y - Xb\ \ '\ P^{-1}\ Y - Xb\ \ = \qquad 2.15$$
$$= \ Y - Xb\ '\ P^{-1}\ '^{P^{-1}}\ Y - Xb\ \ = \ Y - Xb\ '\Omega^{-1}\ Y - Xb\ \ = e'\Omega^{-1}e,$$

т. е. видно, что оценку b^* обобщенного метода наименьших квадратов можно определить как точку минимума обобщенного критерия $e'\Omega^{-1}e$ (2.14).

Для того чтобы коэффициент детерминации являлся состоятельной оценкой качества регрессионной модели вычислим его по формуле:

$$R^2 = \frac{Q_R}{Q} = \frac{b'X'Y' - ny^2}{Y'Y - ny^2},$$

т.е.

$$R^2 = 1 - \frac{(Y - Xb^*)'(Y - Xb^*)}{(Y - Y)'(Y - Y)} \qquad 2.16$$

Качество обобщенной модели показывает коэффициент детерминации R^2. Но данная характеристика является лишь приближенной из-за того что наличие свободного члена в исходной модели не всегда гарантирует его присутствие в полученной модели (2.11). Обычно, значение данного коэффициента выходит даже за пределы интервала [0;1]. Однако, добавление или удаление объясняющих переменных не всегда приводит к увеличению или уменьшению R^2.

Для того чтобы применить обобщенный метод наименьших квадратов необходимо знать ковариационную матрицу вектора возмущений Ω. Данный случай на практике встречается крайне редко.

Если же считать все $n(n+1)/2$ элементов симметричной ковариационной матрицы Ω неизвестными параметрами обобщенной модели, то общее число параметров превышает число наблюдений n, тем самым делает оценку этих параметров неразрешимой задачей. Следовательно, для практической реализации обобщенного МНК необходимо введение дополнительных условий на структуру матрицы Ω.

2.2 Обобщенная линейная модель множественной регрессии при неизвестной ковариационной матрице регрессионных остатков

В большинстве случаев на практике матрица Ω является неизвестной и, как было отмечено в п. 2.1, оценить ее $n(n+1)/2$ параметров по n наблюдениям является невозможным. Пусть задана структура Ω, т. е. форма ее функциональной зависимости от относительно небольшого числа параметров $\theta_1, \theta_2, \dots, \theta_m$, т. е. матрица $_{\varepsilon} = \sigma^2 \Omega_0(\theta_1, \theta_2, \dots, \theta_m)$. Например, в модели с автокоррелированными остатками структура матрицы $_{\varepsilon}$ определется двумя параметрами σ^2 и $\theta_1 = \rho$, матрица Ω_0 выглядит следующим образом:

$$\Omega_0 = \begin{matrix} 1 & \rho & \dots & \rho^{n-1} \\ \rho & 1 & \dots & \rho^{n-2} \\ \dots & \dots & \dots & \dots \\ \rho^{n-1} & \rho^{n-2} & \dots & 1 \end{matrix},$$

ρ — неизвестный параметр, который необходимо оценить.

Для оценки матрицы $_{\varepsilon} = \sigma^2 \Omega_0$ необходимо по исходным наблюдениям найти состоятельные оценки параметров $\theta = (\theta_1, \theta_2, \dots, \theta_n)'$. Затем получить оценку параметра σ^2. Учитывая (2.5) такую оценку для классической модели можно найти разделив минимальную остаточную сумму квадратов $\sum_{i=1}^{n} e_i^2 = e'e$ на число степеней свободы $(n-p-1)$. Применимо к обобщенной модели это будет выглядеть следующим образом:

$$s_*^2 = \frac{e'\Omega_0^{-1}e}{n-p-1} = \frac{(Y-Xb^*)'\Omega_0^{-1}(Y-Xb^*)}{n-p-1}. \qquad 2.17$$

При помощи известной s_*^2, вычислим матрицу $_{\varepsilon} = s_*^2 \Omega_0$.

Доказано, что если использовать полученные оценки вместо неизвестных истинных значений σ^2 и $_{\varepsilon} = \sigma^2 \Omega_0$, то также получим состоятельные оценки параметра β ковариационной матрицы $_{\beta}$. Такой метод носит название доступного (или практически реализуемого) обобщенного метода наименьших квадратов.

Оценка доступного обобщенного МНК β:

$$b^* = (X'\Omega_0^{-1}X)^{-1}X'\Omega_0^{-1}Y. \qquad 2.18$$

При применении метода максимального правдоподобия для оценки нормальной обобщенной линейной модели регрессии оценки максимального правдоподобия β, θ, σ^2 находем из системы уравнений правдоподобия:

$$\beta = (X'\Omega_0^{-1}X)^{-1}X'\Omega_0^{-1}Y, \quad 2.19$$

$$\sigma^2 = \frac{1}{n}e'\Omega_0 e, \quad 2.20$$

$$\frac{e'c_j e}{e'\Omega_0^{-1}e} = \frac{1}{n}tr\ c_j\Omega_0\ , j = 1,\dots,m, \quad 2.21$$

$e = Y - X\beta, c_j = \frac{\partial\Omega_0^{-1}\ \theta}{\partial\theta_j}$.

Проанализировав систему (2.19) — (2.21) можно заметить, что оценки β и σ^2 метода максимального правдоподобия будут совпадать с оценками b^* и s_*^2 обобщенного метода наименьших квадратов.

Решить систему (2.19) — (2.21) можно с помощью итерационной процедуры, например, двухшаговой.

1-й шаг. На первом шаге необходимо найти оценку метода максимального правдоподобия $\beta_0 = (X'X)^{-1}X'Y$.

Вычислим остатки $e_1 = Y - X\beta_1$ и решим полученную систему (2.19) — (2.21) при заданных остатках.

Находем $m \times 1$ вектор θ_1 и матрицу $\Omega_{0(1)} = \Omega_0(\theta_{(1)})$.

2-й шаг. Находим оценку вектора β по формуле (2,18):

$$\beta_2 = (X'\Omega_0^{-1}X)^{-1}X'\Omega_0^{-1}Y.$$

Вычисленные этим методом оценки при большом *n* будут совпадать с оценками рассчитанными методом максимального правдоподобия, следовательно, будут эффективными.

3 ВЫЧИСЛИТЕЛЬНЫЙ ЭКСПЕРИМЕНТ

3.1 Реализация классической линейной модели множественной регрессии

Выполним построение классической регрессионной модели на примере набора Модели номинального ВВП , учитывающая влияние темпов роста индекса потребительских цен (CPI), реальной заработной платы (WR) и денежной массы (MS) [6].

Загрузим данные при помощи следующей функции:

```
load Data_NelsonPlosser
```

Данные показаны на рисунке 2.

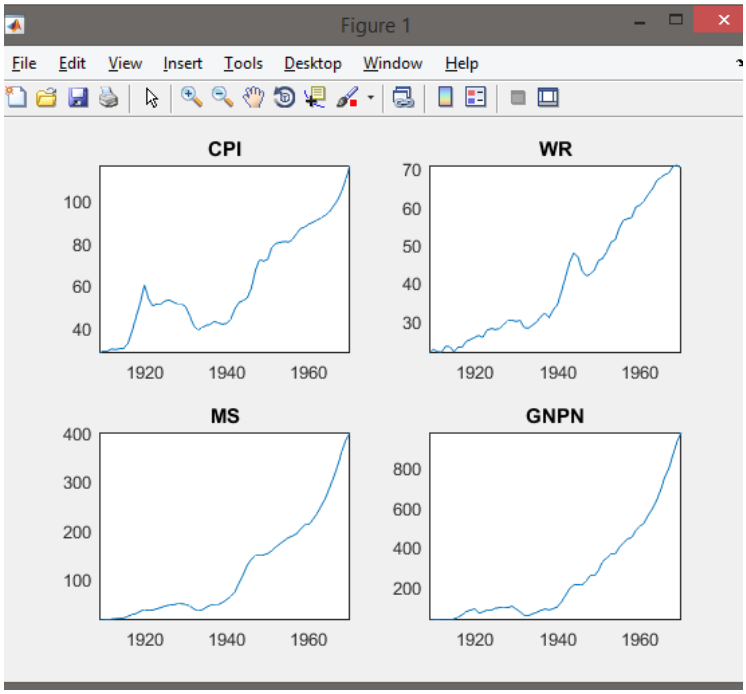

Рисунок 2 – Данные

Предполагая, что все предпосылки регрессионного анализа выполняются, построим регрессионную модель методом наименьших квадратов.

```
Mdl_OLS = fitlm(dLogTbl);
```

Fitlm – функция, которая выполняет построение линейной модели множественной регрессии обычным методом наименьших квадратов для переменных из таблицы или массива данных (ОМНК) [24].

Результат работы данной функции представлен на рисунке 3.

```
Linear regression model:
    GNPNRate ~ 1 + CPIRate + WRRate + MSRate

Estimated Coefficients:
                    Estimate        SE         tStat        pValue
                   __________    _________    ________    __________

    (Intercept)    -0.0076497    0.0084845    -0.90161       0.37106
    CPIRate            0.9037      0.15439      5.8533    2.4978e-07
    WRRate             0.9036      0.19059      4.7411     1.461e-05
    MSRate             0.4285      0.13788      3.1079      0.002938


Number of observations: 61, Error degrees of freedom: 57
Root Mean Squared Error: 0.049
R-squared: 0.764,  Adjusted R-Squared 0.751
F-statistic vs. constant model: 61.4, p-value = 7.42e-18
```

Рисунок 3 – Построение модели ОМНК

Проанализируем полученные данные.

Estimate – коэффициенты регрессионной модели, полученные методом наименьших квадратов, показывают силу и тип связи объясняющих переменных с зависимой. Т.к. модель линейная, мы получили оценки параметров, при которых сумма квадратов отклонений фактических значений результативного признака от теоретических минимальна. В общем случае коэффициент регрессии показывает, как в среднем изменится результативный признак Y, если факторный признак X увеличится на единицу.

$$Y = -0.00765 + 0.9037X_1 + 0.9036X_2 + 0.4285X_3$$

В данном случае при росте индекса потребительских цен (CPI) на 1 ед. объем номинального ВВП вырастет на 0,9037 ед., аналогично при росте

реальной заработной платы (WR) и денежной массы (MS) объем ВВП вырастет на 0,9036 и 0,4285 ед. соответственно.

Intercept – (y-пересечение) – коэффициент, который показывает каким будет Y в случае, если все используемые в модели факторы будет равны, подразумевается, что это зависимость от других, неописанных в модели факторов.

R-squared – коэффициент детерминации (R^2), равный 0,764. Показывает, что на 76,4% расчетные параметры модели, т.е. сама модель, объясняют зависимость и изменения изучаемого параметра – Y от исследуемых факторов X. Можно сказать, что это показатель качества модели, и чем он выше, тем лучше. Очевидно, что коэффициент детерминации не может быть больше 1. Чем ближе значение коэффициента к 1, тем сильнее зависимость, т.е. данные наиболее соответствуют модели. Модели с коэффициентов более 80% можно признать достаточно хорошими, но если коэффициент менее 50%, то такую модель можно смело ставить под сомнение. Равенство данного коэффициента единице показывает, что зависимая переменная в точности описывается полученной моделью.

Adjusted R-squared – скорректированный R^2. С его помощью можно сравнивать модели с разным числом признаков так, чтобы их число не влияло на статистику R^2. Для моделей с одинаковой зависимой переменной и одинаковым объемом выборки сравнение моделей с помощью скорректированного коэффициента детерминации эквивалентно их сравнению с помощью остаточной дисперсии или стандартной ошибки модели.

F-статистика, называемая также критерием Фишера, используется для оценки значимости линейной зависимости, опровергая или подтверждая гипотезу о ее существовании. Значение t-статистики (критерий Стьюдента) помогает оценивать значимость коэффициента при неизвестной либо

свободного члена линейной зависимости. Если значение t-критерия > tкр, то гипотеза о незначимости свободного члена линейного уравнения отвергается.

Проверим значимость полученного уравнения регрессии. Фактическое значение критерия F = 61,4 больше табличного $F_{0,05;3;57} = 2,76$ определенного на уровне значимости α=0,05, при $k_1 = 3$, $k_2 = 61 - 3 - 1 = 57$ степенях свободы, т. е. уравнение регрессии значимо, следовательно, исследуемая зависимая переменная Y достаточно хорошо описывается включенными в регрессионную модель переменными X_1, X_2, X_3.

3.2 Реализация обобщенной линейной модели множественной регрессии

На графике видно, что данные достаточно неоднородны, что приведет к различной дисперсии ошибок, т.е. нарушению предпосылки 3 регрессионного анализа о гомоскедастичности. Гетероскедастичность является частным случаем нарушения этой предпосылки. Следовательно, оценки метода наименьших квадратов окажутся неэффективными, а формулы для вычисления дисперсий коэффициентов и статистик, которые используются для проверки гипотез – неверными. Устранить данную проблему можно двумя способами: применив доступный обобщенный метод наименьших квадратов или перейти от значений переменных, например к их первым производным, взятым от логарифма переменной.

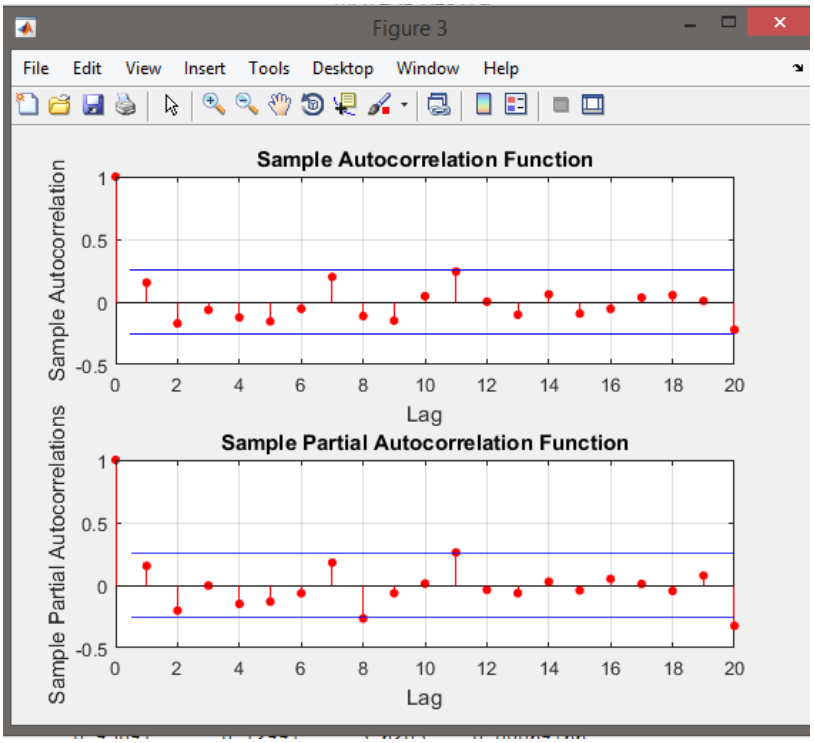

Рисунок 4 – Автокорреляция

Воспользуемся вторым способом: прологарифмируем и затем продифференцируем значения всех переменных, чтобы получить однородные данные.

```
dLogTbl = array2table(diff(log(Tbl{:,:})),...
    'VariableNames',strcat(Tbl.Properties.VariableNames,'Rate'));
```

Выполним построение обобщенной регрессионной модели на примере данных, приведенных в предыдущем пункте, с помощью функции fgls.

```
[coeff,se,EstCoeffCov] = fgls(dLogTbl,'innovMdl','HC0','display','final');
```

Fgls – функция, которая рассчитывает точечные оценки коэффициентов обобщенного линейного уравнения регрессии b доступным обобщенным методом наименьших квадратов (ДОМНК) [25].

Полученные результаты представлены на рисунке 5.

```
Linear regression model:
    GNPNRate ~ 1 + CPIRate*WRRate + CPIRate*MSRate

Estimated Coefficients:
                       Estimate        SE         tStat       pValue
                       ________    _________    _______    __________

    (Intercept)        0.0063657    0.0083147     0.7656       0.44719
    CPIRate               1.1227      0.15371     7.3038     1.1988e-09
    WRRate                0.5768      0.20288     2.8431      0.0062589
    MSRate               0.45041      0.12441     3.6203     0.00064166
    CPIRate:WRRate        8.0755       3.0028     2.6893      0.0094587
    CPIRate:MSRate       -5.5406       1.3431    -4.1251     0.00012639


Number of observations: 61, Error degrees of freedom: 55
Root Mean Squared Error: 0.0434
R-squared: 0.82,  Adjusted R-Squared 0.804
F-statistic vs. constant model: 50.3, p-value = 2.69e-19
>>
```

Рисунок 5 – Модель ДОМНК

Модель будет выглядеть следующим образом:

$$Y = 0.0064 + 1.1227X_1 + 0.5768X_2 + 0.45041X_3.$$

В данном случае при росте индекса потребительских цен (CPI) на 1 ед. объем номинального ВВП вырастет на 1.1227 ед., аналогично при росте реальной заработной платы (WR) и денежной массы (MS) объем ВВП вырастет на 0.5768 и 0.45041 ед. соответственно.

R-squared – коэффициент детерминации (R^2), равный 0,82. Таким образом, видно, что данная модель на 82% объясняет зависимость и изменения изучаемого параметра – Y от исследуемых факторов X.

Проверим значимость полученного уравнения регрессии. Фактическое значение критерия F = 50,3 больше табличного $F_{0,05;3;57} = 2,76$ определенного на уровне значимости α=0,05, при $k_1 = 3$, $k_2 = 61 - 3 - 1 = 57$ степенях свободы, т. е. уравнение регрессии значимо, следовательно, исследуемая зависимая переменная Y достаточно хорошо описывается включенными в регрессионную модель переменными X_1, X_2, X_3.

3.3 Сравнительный анализ классического и обобщенного метода наименьших квадратов

Построенные уравнения регрессии редко удовлетворяют необходимым характеристикам. Поэтому необходимо оценить результаты моделирования.

О корректности модели могут сказать следующие характеристики:

1. Стандартная ошибка уравнения регрессии;
2. Общее качество уравнения регрессии;
3. Стандартная ошибка параметров уравнения;
4. Выполнимость предпосылок МНК:
    - оценка автокорреляции остатка;
    - оценка мультиколлинеарности переменных;
5. корректность модели в целом [9].

Стандартная ошибка с правильной степенью свободы может быть рассчитана следующим образом:

$$S_y = \sqrt{\frac{y - y^{\,2}}{(n-1)}}$$

Ошибка рассчитанная данным образом будет характеризовать абсолютную величину разброса случайной составляющей регрессионного уравнения [14].

В программном пакете часто используется одна процедура для того чтобы обхватить все вышеперечисленные модели.

Но есть некоторые ограничения GLM, например, линейная функция может иметь только линейный предиктор в систематической компоненте, оценки должны быть независимыми.

Чтобы определить практическую значимость уравнения множественной регрессии необходимо вычислить показатель корреляции и детерминации. Показатель корреляции показывает насколько тесно связаны зависимая и объясняющая переменные, а также – совместное влияние факторов на результат.

Коэффициент корреляции имеет несколько преимуществ по сравнению с ковариацией для определения сильных сторон отношений:

- Ковариация может принимать практически любое число, в то время как корреляция ограничена: от -1 до +1.
- Из-за его числовых ограничений корреляция более полезна для определения того, насколько сильная взаимосвязь между этими двумя переменными.
- Корреляция не имеет единиц. Ковариация всегда имеет единицы
- На корреляцию не влияют изменения в центре (т. Е. Среднее значение) или масштаб переменных.

Дисперсионный анализ (ANOVA) - это инструмент анализа, используемый в статистике, который разбивает совокупную изменчивость, найденную внутри набора данных, на две части: систематические факторы и случайные факторы. Систематические факторы оказывают статистическое влияние на данный набор данных, а случайные - нет. Аналитики используют анализ дисперсионного теста для определения результата, который независимые переменные имеют на зависимой переменной на фоне регрессионного исследования [22].

Анализ теста на отклонения является начальным этапом анализа факторов, которые влияют на данный набор данных. Как только анализ теста на отклонение завершен, аналитик проводит дополнительное тестирование на методические факторы, которые в значительной степени способствуют несогласованности набора данных. Аналитик использует анализ результатов теста дисперсии в f-тесте для генерации дополнительных данных, которые согласуются с предлагаемыми регрессионными моделями [26].

Тест позволяет одновременно сравнивать более двух групп, чтобы определить, существует ли между ними взаимосвязь. Тест анализирует несколько групп для определения типов между образцами и внутри них. Например, исследователь может опробовать студентов из нескольких

колледжей, чтобы убедиться, что студенты из одного из колледжей превзошли других. Кроме того, исследователь R & D может протестировать два разных процесса создания продукта, чтобы убедиться, что один процесс лучше, чем другой, с точки зрения эффективности затрат [27].

Тип запуска ANOVA зависит от ряда факторов. Он применяется, когда данные должны быть экспериментальными. Анализ дисперсии используется, если нет доступа к статистическому программному обеспечению, что приводит к вычислению ANOVA вручную. Он прост в использовании и наилучшим образом подходит для небольших образцов. Со многими экспериментальными проектами размеры выборки должны быть одинаковыми для разных комбинаций факторов.

Анализ дисперсий полезен для тестирования трех или более переменных. Это похоже на множественные t-тесты с двумя образцами. Однако это приводит к меньшему количеству ошибок типа I и подходит для решения ряда проблем. ANOVA группирует различия, сравнивая средства каждой группы и включает распространение дисперсии на различные источники. Он используется с предметами, группами тестирования, между группами и внутри групп.

Существует два типа дисперсионного анализа: односторонний (или однонаправленный) и двухсторонний. Односторонний или двусторонний относится к числу независимых переменных в вашем тесте Analysis of Variance. Односторонний ANOVA оценивает влияние единственного фактора на единственную переменную ответа. Он определяет, все ли образцы одинаковы. Односторонний ANOVA используется для определения наличия статистически значимых различий между средствами трех или более независимых (несвязанных) групп [15].

Двухсторонний ANOVA является расширением одностороннего ANOVA. В одностороннем порядке у вас есть одна независимая переменная, влияющая на зависимую переменную. С двухсторонним ANOVA существует два независимых. Например, двухсторонний ANOVA позволяет компании

сравнивать производительность труда на основе двух независимых переменных, например, оклада и набор навыков. Он используется для наблюдения за взаимодействием между двумя факторами. Он одновременно проверяет влияние двух факторов [13].

T- и z-тесты, разработанные в 20-м веке, использовались до 1918 года, когда Рональд Фишер создал анализ дисперсии. ANOVA также называют дисперсионным анализом Фишера и является расширением t- и z-тестов. Этот термин стал известен в 1925 году, появившись в книге Фишера «Статистические методы для исследователей». Он использовался в экспериментальной психологии, а затем расширился до более сложных предметов [16].

Формула для F, используемая в ANOVA, равна F = между оценкой групповой дисперсии (MSB), деленной на оценку групповой дисперсии (MSW), где F = MSB / MSW. Каждая оценка дисперсии состоит из двух частей: суммы квадратов и обода (SSB и SSW) и степеней свободы (df).

ЗАКЛЮЧЕНИЕ

В данной выпускной квалификационной работе были поставлены следующие задачи:

1. Проанализировать алгоритм классической линейной модели множественной регрессии.
2. Проанализировать регрессионную модель в условиях нарушения классических предпосылок.
3. Осуществить анализ реализаций линейной множественной регрессионной модели.

В процессе работы были реализованы и проаналиированы классическая и обобщенная регрессионные модели.

СПИСОК ИСПОЛЬЗУЕМОЙ ЛИТЕРАТУРЫ

1. Estimasi Model Seemingly Unrelated Regression (SUR) dengan Metode Generalized Least Square (GLS), Ade Widyaningsih, Made Susilawati, I Wayan Sumarjaya, Jurnal Matematika]. Режим доступа: https://ojs.unud.ac.id/index.php/jmat/article/view/12554 – Дата обновления: 01апреля 2018 г.

2. Generalized Linear Models [Электронный ресурс]. Режим доступа: http://scikit-learn.org/stable/modules/linear_model.html – Дата обновления: 01апреля 2018 г.

3. Generalized Linear Models [Электронный ресурс]. Режим доступа: http://isu.ifmo.ru/docs/doc112/datamine.112/e12216/algo_glm.htm – Дата обновления: 01апреля 2018 г.

4. Generalized Linear Models [Электронный ресурс]. Режим доступа: https://www.ibm.com/support/knowledgecenter/SSLVMB_23.0.0/ spss/advanced/idh_idd_genlin_typeofmodel.html – Дата обновления: 01апреля 2018 г.

5. Introduction to Generalized Linear Models [Электронный ресурс]. Режим доступа: https://onlinecourses.science.psu.edu/stat504/node/216 – Дата обновления: 01апреля 2018 г.

6. MATLAB – высокоуровневый язык технических расчётов [Электронный ресурс] Режим доступа: https://matlab.ru/products/matlab. – Дата обновления: 20 ноября 2017г.

7. Агалаков С.А., Эконометрические модели [Текст] : учебное пособие / С. А. Агалаков ; М-во образования и науки Российской Федерации, Федеральное гос. бюджетное образовательное учреждение высш. проф. образования Омский гос. ун-т им. Ф. М. Достоевского. - Омск : Изд-во Омского гос. ун-та им. Ф. М. Достоевского, 2015. - 140 с. : ил., табл.; 20 см.; ISBN 978-5-7779-1820-8.

8. Буре, В.М. Методы прикладной статистики в R и Excel [Электронный ресурс] : учебное пособие / В.М. Буре, Е.М. Парилина, А.А. Седаков. — Электрон. дан. — Санкт-Петербург : Лань, 2018. — 152 с. — Режим доступа: https://e.lanbook.com/book/104938.

9. Гладилин А.В., Эконометрика [Текст]: учебное пособие для студентов высших учебных заведений, обучающихся по экономическим специальностям / А. В. Гладилин, А. Н. Герасимов, Е. И. Громов. - 3-е изд., стер. - Москва : КноРус, 2011. - 227 с. : ил., табл.; 21 см.; ISBN 978-5-406-00943-7.

10. Деркаченко, В.Н. Эконометрика [Электронный ресурс] / В.Н. Деркаченко. — Электрон. дан. — Пенза : ПензГТУ, 2013. — 140 с. — Режим доступа: https://e.lanbook.com/book/62724.

11. Домбровский В.В., Эконометрика [Текст]: учебник / В.В. Домбровский; Федер. агентство по образованию, Нац. фонд подгот. кадров. – М.: Новый учебник, 2004. – 342 с., – ISBN 5-8393-0400-X.

12. Доступная обобщенная регрессионная модель [Электронный ресурс] Режим доступа: https://www.mathworks.com/help/econ/fgls.html. – Дата обновления: 20 ноября 2017г.

13. Дьяков, И.И. Основы эконометрики [Электронный ресурс] : учебное пособие / И.И. Дьяков, И.В. Жуплей. — Электрон. дан. — Уссурийск : Приморская ГСХА, 2013. — 103 с. — Режим доступа: https://e.lanbook.com/book/69558.

14. Елисеева И.И. - Отв. ред., Эконометрика : учебник для бакалавриата и магистратуры / И. И. Елисеева [и др.] ; под ред. И. И. Елисеевой. — М. : Издательство Юрайт, 2015. — 449 с. — (Серия : Бакалавр и магистр. Академический курс). — ISBN 978-5-9916-5161-5.

15. Заяц, О.А. Эконометрика [Электронный ресурс] : учебное пособие / О.А. Заяц. — Электрон. дан. — Волгоград : Волгоградский ГАУ, 2016. — 96 с. — Режим доступа: https://e.lanbook.com/book/76670.

16. Картаев, Ф.С. Эконометрика [Электронный ресурс] : учебное пособие / Ф.С. Картаев, Е.Н. Лукаш. — Электрон. дан. — Москва : Экономический факультет МГУ им. М.В. Ломоносова, 2014. — 118 с. — Режим доступа: https://e.lanbook.com/book/73154.

17. Классическая линейная регрессионная модель [Электронный ресурс] Режим доступа: https://www.mathworks.com/help/stats/regress.html?searchHighlight=regress&s_tid=doc_srchtitle – Дата обновления: 20 ноября 2017 г.

18. Кремер Н.Ш., Путко Б.А., Эконометрика [Текст]: Учебник для вузов / Под ред. проф. Н.Ш. Кремера. – М.: ЮНИТИ-ДАНА, 2003. – 311 с., – ISBN 5-238-00333-1.

19. Литвинова, И.А. Эконометрика [Электронный ресурс] : учебное пособие / И.А. Литвинова. — Электрон. дан. — Кемерово : КемГУ, 2017. — 103 с. — Режим доступа: https://e.lanbook.com/book/102673.

20. Магнус Я.Р., Катышев П.К., Пересецкий А.А., Эконометрика. Начальный курс: Учеб. – 6-е изд., перераб. и доп. – М.: Дело, 2004. – 576 с., – ISBN 5-7749-0055-X.

21. Мхитарян В.С., Архипова М.Ю., Балаш В.А., Эконометрика [Текст]: учебник / Под ред. В. С. Мхитаряна. - Москва : Проспект, 2014. - 380 с. : ил., табл.; ISBN 978-5-392-13469-4.

22. Новиков, А.И. Эконометрика [Электронный ресурс] : учебное пособие / А.И. Новиков. — Электрон. дан. — Москва : Дашков и К, 2017. — 224 с. — Режим доступа: https://e.lanbook.com/book/93399.

23. Приходько А. И., Практикум по эконометрике. Регрессионный анализ средствами Excel : Учеб. пособие для студентов вузов, обучающихся по спец. 061800 "Математические методы в экономике"... / Ин-т экономики, права и естеств. спец. и др. - Краснодар, 2004 (Тип. Экоинвест). - 179 с.; ISBN 5-9488-017-6 (в обл.).

24. Ревинская О. Г. Р32 Основы программирования в MatLab: учеб. пособие. — СПб.: БХВ-Петербург, 2016. — 208 с.: ил. — (Учебное пособие) ISBN 978-5-9775-3564-9.

25. Список функций Statistics Toolbox [Электронный ресурс]. Режим доступа: http://matlab.exponenta.ru/statist/book2/11/regress.php. – Дата обновления: 01апреля 2018 г.

26. Эконометрика: продвинутый курс с приложениями в финансах [Текст] : учебник / С. А. Айвазян, Д. Фантаццини ; Московская школа экономики МГУ им. М. В. Ломоносова. - Москва : Магистр : ИНФРА-М, 2014. - 942 с. : ил., табл.; 25 см.; ISBN 978-5-9776-0333-1.

27. Яновский, Л.П. Введение в эконометрику [Электронный ресурс] : учебное пособие / Л.П. Яновский, А.Г. Буховец. — Электрон. дан. — Москва: КноРус, 2015. — 256 с. — Режим доступа: https://e.lanbook.com/book/53398.