Министерство образования и науки Российской Федерации

Федеральное государственное бюджетное
образовательное учреждение
высшего профессионального образования
«Оренбургский государственный университет»

Кафедра математических методов и моделей в экономике

ПОСТРОЕНИЕ И ИССЛЕДОВАНИЕ ЛИНЕЙНОЙ МОДЕЛИ МНОЖЕСТВЕННОЙ РЕГРЕССИИ В УСЛОВИЯХ ПЛОХОЙ ОБУСЛОВЛЕННОСТИ НОРМАЛЬНОЙ СИСТЕМЫ ЛИНЕЙНЫХ УРАВНЕНИЙ

Под редакцией А.Г. Реннера

Рекомендовано к изданию Редакционно-издательским советом федерального государственного бюджетного образовательного учреждения высшего профессионального образования «Оренбургский государственный университет» в качестве методических указаний для студентов, обучающихся по программам высшего профессионального образования по специальности 080016.65 Математические методы в экономике, направлениям подготовки 231300.62 Прикладная математика «Общий профиль», 080500.62 Бизнес-информатика профиль «Архитектура предприятия», 080100.62 Экономика «Общий профиль», профиль «Математические методы в экономике»

Оренбург
2012

УДК 330.4(076)

ББК 65В631Я7
П63

Рецензент – доцент, кандидат экономических наук Т.В. Леушина

Авторы: О.И. Бантикова, В.И. Васянина, Ю.А. Жемчужникова, О.Н. Яркова

П 67 Построение и исследование линейной модели множественной регрессии в условиях плохой обусловленности нормальной системы линейных уравнений: методические указания /О.И. Бантикова, В.И. Васянина, Ю.А. Жемчужникова, О.Н. Яркова; под ред. А.Г. Реннера; Оренбургский гос. ун-т.– Оренбург: ОГУ, 2012. – 40 с.

Методические указания к семинарским занятиям, лабораторному практикуму, самостоятельной работе студентов, в том числе для выполнения расчетно-графических заданий, курсовых и дипломных работ, предполагающих проведение регрессионного анализа.

Предназначены для студентов специальности 08001.65 Математические методы в экономике, направлений подготовки 231300.62 Прикладная математика «Общий профиль», 080500.62 Бизнес-информатика профиль «Архитектура предприятия», 080100.62 Экономика «Общий профиль», профиль «Математические методы в экономике» и других специальностей и направлений.

УДК 330.4(076)

ББК 65В631Я7

Содержание

Введение

При построении МНК-оценок линейной модели множественной регрессии исследователь зачастую сталкивается с проблемой плохой обусловленности нормальной системы линейных уравнений, которая помимо того, что влечет за собой погрешности в вычислении МНК-оценок коэффициентов и неустойчивость оценок к незначительным изменениям исходных данных, может привести к неверным статистическим выводам относительно значимости модели и значимости отдельных коэффициентов.

Цель методических указаний – способствовать приобретению практических навыков исследования линейной модели множественной регрессии в условиях плохой обусловленности нормальной системы линейных уравнений.

1 Теоретическая часть

1.1 Общая постановка задачи регрессионного анализа

Ставится задача построения и исследования регрессионной зависимости результирующего признака *y* от объясняющих переменных $x_0 = 1, x_1, x_2, ..., x_k$ на основе результатов наблюдений признаков на "n" объектах $O_1, O_2, ...O_n$, $n >> k$.

Результаты наблюдений результирующего признака и объясняющих переменных представлены вектором $Y_{n\times 1} = (y_1 \quad y_2 \quad ... \quad y_n)^T$ и матрицей X типа «объект-свойство»:

$$X_{n\times k+1} = \begin{pmatrix} 1 & x_{11} & x_{12} & ... & x_{1k} \\ 1 & x_{21} & x_{22} & ... & x_{2k} \\ ... & ... & ... & ... & ... \\ 1 & x_{n1} & x_{n2} & ... & x_{nk} \end{pmatrix} = \{x_{ij}\}_{\substack{i=\overline{1,n} \\ j=\overline{0,k}}}$$

где y_i – наблюденное значение результативного признака для i-го объекта;

x_{ij} – значение j-го признака на i-м объекте наблюдения $i = \overline{1,n}$, $j = \overline{0,k}$; столбец из "1" можно считать столбцом "наблюденных" значений для признака $x_0 = 1$.

Регрессионную зависимость результативной переменной y от объясняющих переменных $x = (x_1, x_2, ..., x_k)^T$ будем искать в виде:

$$\tilde{y} = \beta_0\psi_0(x) + \beta_1\psi_1(x) + ... + \beta_k\psi_k(x), \qquad (1.1)$$

где $\tilde{y}$ – условное среднее значение результативной переменной y для каждого фиксированного набора значений объясняющих переменных;

$\psi_i(x), i = \overline{0k}$ - линейно независимые базисные функции;

$\psi_0(x) \equiv 1$;

$\beta = (\beta_0 \ \beta_1 ... \beta_k)^T$ - вектор коэффициентов функции регрессии.

“Линейная” модель регрессии (линейная по неизвестным коэффициентам) будет иметь вид:

$$y_i = \beta_0\psi_0(x_{i1}x_{i2...}x_{ik}) + \beta_1\psi_1(x_{i1}x_{i2...}x_{ik}) + ...\beta_k\psi_k(x_{i1}x_{i2...}x_{ik}) + z_i = y_i = \beta^T\psi(x_i) + z_i \ , \ i = \overline{1,n} \ , \qquad (1.2)$$

где y_i - наблюденное значение результативного признака *i*–го объекта;

$\psi = (\psi_0 \quad \psi_1 \quad .... \quad \psi_k)^T$;

x_i - наблюденные значения объясняющих переменных *i*-ого объекта наблюдения;

z_i - значение регрессионного остатка, характеризующего отклонения наблюденных значений y_i от модельных значений $\tilde{y}_i$ для *i*–го объекта.

В матрично-векторных обозначениях выражение (1.2) примет вид:

$$Y = \Psi\beta + Z \ , \qquad (1.2a)$$

где $\Psi_{n\times(k+1)} = \begin{pmatrix} \psi_0(x_{11}x_{12}...x_{1k}) & \psi_1(x_{11}x_{12}...x_{1k}) & .... & \psi_k(x_{11}x_{12}...x_{1k}) \\ \psi_0(x_{21}x_{22}...x_{2k}) & \psi_1(x_{21}x_{22}...x_{2k}) & .... & \psi_k(x_{21}x_{22}...x_{2k}) \\ .... & .... & ... & .... \\ \psi_0(x_{n1}x_{n2}...x_{nk}) & \psi_1(x_{n1}x_{n2}...x_{nk}) & ... & \psi_k(x_{n1}x_{n2}...x_{nk}) \end{pmatrix}$,

$Z = (z_1,...,z_n)^T$ – вектор значений регрессионных остатков.

На $Z = (z_1,...,z_n)^T$ – можно смотреть как на возможные значения случайной величины $\varepsilon = (\varepsilon_1,...,\varepsilon_n)^T$, на Y - как на возможные значения случайной величины $\eta_{1,n}$, тогда выборочная модель будет иметь вид:

$$\eta_{1,n} = \Psi\beta + \varepsilon \ , \qquad (1.3)$$

где $\eta_{1,n} = (\eta_1, \eta_2, ..., \eta_n)^T$ - случайный вектор, а (1.2a) – реализация модели (1.3).

Для нахождения оценок реализации выборочной модели множественной регрессии (1.2а) имеем систему нормальных уравнений (1.4):

$$\Psi^T \Psi \beta = \Psi^T Y . \qquad (1.4)$$

Решением системы (1.4), в случае выполнения второго условия Гаусса – Маркова, является вектор МНК-оценок, формально записанный в виде (1.5):

$$\hat{\beta}_{\text{МНК}} = (\Psi^T \Psi)^{-1} \Psi^T Y \qquad (1.5)$$

Фактически система нормальных уравнений решается, в зависимости от ее свойств, одним из методов линейной алгебры.

1.2 Проблема плохой обусловленности МНК-оценок ЛММР

Определение. Система линейных уравнений, например (1.4), называется *плохо обусловленной*, если малые возмущения элементов правой части системы уравнений (1.4) или матрицы $(\Psi^T \Psi)$ или того и другого вместе приводит к большим изменениям в решении этой системы.

Признаки плохой обусловленности

1. Среди коэффициентов уравнения регрессии много, а может быть и все незначимы, а модель в целом является значимой.

2. Стандартные отклонения велики настолько, что сравнимы или даже превосходят сами коэффициенты.

3. Доверительные интервалы для коэффициентов уравнения регрессии содержат внутри себя точку нуль.

4. Малость определителя матрицы $\Psi^T \Psi$.

Признаки 1-4 являются необходимыми условиями плохой обусловленности.

5. Достаточным условием плохой обусловленности является большое значение числа обусловленности (порядка сотни и более).

$$M = \frac{\max\limits_{i=1n}|\lambda_i|}{\min\limits_{i=1n}|\lambda_i|}, \qquad (1.6)$$

где λ_i - собственное число матрицы $\Psi^T\Psi$.

Если система линейных уравнений является плохо обусловленной, то решение системы (1.4) стандартными методами повлечет за собой погрешности в вычислении МНК-оценок коэффициентов, неверные статистические выводы относительно значимости модели и значимости отдельных коэффициентов.

Для устранения этих проблем необходимо поиск решений системы (1.4) осуществлять методом регуляризации или рекуррентным методом наименьших квадратов.

1.2.1 Метод регуляризации

Предположим, что система линейных алгебраических уравнений, например

$$A\alpha = B, \qquad (1.7)$$

где A - матрица коэффициентов системы;

B – вектор правых частей;

α - вектор неизвестных,

является плохо обусловленной. Перепишем (1.7) в эквивалентном виде

$$\langle A\alpha - B, A\alpha - B\rangle = 0, \qquad (1.8)$$

где $\langle \bullet,\bullet \rangle$ - скалярное произведение векторов.

От задачи (1.8) перейдем к регуляризованной системе линейных алгебраических уравнений:

$$\left(A^T A + \tau E\right)\alpha = A^T B + \tau \alpha^0 \qquad (1.9)$$

где E – единичная матрица;

τ - параметр регуляризации ($0 < \tau < 1$);

α^0 - заданный начальный вектор.

Выбор параметра регуляризации зависит от требуемой точности решения и требований к устойчивости вычислительных алгоритмов. Очевидно, что, чем меньше τ, тем хуже обусловленность матрицы системы (1.9), что приводит к неустойчивости численного решения. При больших значениях τ система (1.9) переходит в хорошо обусловленную систему и ее решение будет сильно отличаться от решения исходной системы (1.7). Поэтому параметр регуляризации выбирают таким образом, что бы система (1.9) была удовлетворительно обусловлена (число обусловленности матрицы $A^T A + \tau E$ несколько десятков) и вместе с тем ее решение не сильно отличалось от решения исходной системы. Выбор такого значения τ является основной проблемой, возникающей при использовании рассматриваемого способа регуляризации плохо обусловленных систем линейных алгебраических уравнений. На практике, для определения подходящего значения параметра регуляризации пользуются невязкой вида:

$$r_\tau = A\alpha_\tau - B\,, \qquad (1.10)$$

где α_τ - решение системы (1.9) при фиксированном значении τ.

Далее возможны два способа выбора параметра регуляризации.

1 способ. Эту невязку сравнивают по норме с известной погрешностью исходных данных. Если τ очень велико, то r_τ по норме много больше этих погрешностей. Если τ мало, то r_τ по норме намного меньше этих погрешностей. Поэтому проводят серию расчетов и в качестве оптимального τ выбирают то, при котором выполняется условие:

$$\|r_\tau\| \approx \|\Delta B\| + \|\Delta A\|,$$

где $\|\bullet\|$ - норма вектора/матрицы;

ΔB -известная погрешность правых частей;

ΔA -известная погрешность матрицы коэффициентов.

2 способ. В качестве оптимального выбирают такое значение τ, при котором $|r_\tau|$ принимает минимальное значение (квазиоптимальный метод выбора параметра регуляризации).

Система (1.9) может быть решена любым из методов решения систем линейных алгебраических уравнений, например, методом Гаусса или с учетом симметричности матрицы коэффициентов методом квадратных корней.

В качестве α^0 можно взять нулевой вектор, тогда система (1.9) примет вид

$$\left(A^T A + \tau E\right)\alpha = A^T B. \qquad (1.11)$$

1.2.2 Рекуррентный метод наименьших квадратов (РМНК)

Предположим, что поиск МНК-оценки осуществляется не по всему массиву $y_1, y_2, \ldots y_n$ экспериментальных данных, а лишь по части $y_1, y_2, \ldots y_m$, $m<n$.

Рекуррентным методом наименьших квадратов (РМНК) называется совокупность выражений (1.12), (1.13):

$$R_m = R_{m-1} - R_{m-1}\psi(x^{(m)})(\psi^T(x^{(m)}R_{m-1}\psi(x^{(m)})+1)^{-1}\psi^T(x^{(m)})R_{m-1}\ , \qquad m=\overline{1n} \qquad (1.12)$$

$$\hat{\beta}^{(m)} = \hat{\beta}^{(m-1)} + R_m\psi(x^{(m)})(y_m - \psi^T(x^{(m)})\hat{\beta}^{(m-1)})\,, \qquad m=\overline{1n} \qquad (1.13)$$

где $\hat{\beta}^{(m)} = \left(\hat{\beta}_0^{(m)} \quad \hat{\beta}_1^{(m)} \quad ... \quad \hat{\beta}_k^{(m)}\right)^T$ - вектор коэффициентов, оцененных по m объектам;

$\hat{\beta}^{(m-1)} = \left(\hat{\beta}_0^{(m-1)} \quad \hat{\beta}_1^{(m-1)} \quad ... \quad \hat{\beta}_k^{(m-1)}\right)^T$ - вектор коэффициентов, оцененных по *m-1* объектам;

$\psi = \left(\psi_0 \quad \psi_1 \quad .... \quad \psi_k\right)^T$

$x^{(m)}$- наблюденные значения объясняющих переменных m-ого объекта наблюдения;

y_m - наблюденное значение результативного признака m-ого объекта наблюдения.

Алгоритм рекуррентного метода наименьших квадратов включает следующие этапы:

1. Задают начальные условия, наиболее часто в виде $\hat{\beta}^{(0)} = 0$, $R_0 = \gamma E$, где $\gamma = cons >> 1$;

2. Полагают m=1 и из условия (1.12) находят R_1, а из (1.13) - $\hat{\beta}^{(1)}$, что формально соответствует поиску оценки $\hat{\beta}^{(1)}$регрессионных параметров по данным для первого объекта наблюдения.

3. Принимают m=2, из условия (1.12) находят R_2, а из (1.13) - $\hat{\beta}^{(2)}$, что соответствует уточненной по данным для первых двух объектов наблюдений.

4. Аналогичным образом проводятся вычисления при m=3,4…n, что приводит к оценке $\hat{\beta}^{(n)}$, принимаемой за МНК-оценку.

Отметим, что никакого серьезного внимания к оценкам $\hat{\beta}^{(1)}$, $\hat{\beta}^{(2)}$, $\hat{\beta}^{(3)}$ проявлять нельзя, они лишены какого-либо практического смысла и должны рассматриваться как формальный “эпизод” на пути получения МНК-оценки.

К достоинствам рекуррентного метода наименьших квадратов следует отнести то, что, во-первых, не приходится обращать матрицу, а, следовательно, не возникает проблем, связанных с обращением плохо обусловленной матрицы, что способствует получению устойчивого решения. Во-вторых, рекуррентный подход позволяет, по мере поступления новых данных, обрабатывать информацию последовательно. На каждом шаге рекуррентных вычислений полученные на предыдущем шаге оценки обновляются с учетом новой порции данных. В ходе их выполнения будут получены оценки параметров для промежуточных моментов. Если в качестве объектов наблюдения выступает время, то полученные оценки для промежуточных моментов позволят проследить динамику коэффициентов уравнения регрессии.

1.3 Мультиколлинеарность: понятие, признаки и методы устранения

Рассмотрим частный случай: $\psi_0(x) = 1, \psi_1(x) = x_1, ...\psi_k(x) = x_k$.

Если между объясняющими переменными существует корреляционная связь, то следует ожидать плохую обусловленность нормальной системы линейных уравнений. В этом случае ее называют мультиколлинеарностью.

Если между объясняющими переменными существует линейная функциональная связь (т.е. значения по меньшей мере одной из них могут быть выражены в виде линейной комбинации наблюденных значений остальных переменных), то rgX<к+1, матрица $X^T X$ -вырожденная (определитель равен нулю), следовательно не существует обратной матрицы, и мы не сможем оценить коэффициенты уравнения регрессии методом наименьших квадратов. Это явление называется **полной мультиколлениарностью.**

Полная мультиколлинеарность встречается достаточно редко, так как ее несложно избежать на предварительной стадии анализа и отбора множества объясняющих переменных путем исключения дублирующих признаков. Формально для выявления полной мультиколлинеарности определяется ранг X (например, методом элементарных преобразований) и попутно выявляются, какие столбцы

линейно зависят от других. Выявив эти столбцы, из модели линейной регрессии исключаются соответствующие этим столбцам признаки, и строится регрессионная модель меньшей размерности по линейно независимым признакам, при этом матрица $X^T X$ будет невырожденная, и, следовательно, существует возможность построения регрессионной модели

Реальная (или частичная) мультиколлинеарность возникает в случаях существования достаточно тесных линейных корреляционных связей между объясняющими переменными.

При сильной корреляционной связи объясняющих переменных, определитель матрицы $X^T X$ будет близок к нулю. Элементы обратной матрицы $(X^T X)^{-1}$ вычисляются с большой погрешностью, следовательно, и оценки, полученные МНК, тоже определяются с погрешностью. Одновременно близость определителя матрицы $(X^T X)$ к нулю влечет за собой большие значения диагональных элементов ковариационной матрицы вектора оценок $\hat{\Sigma}_{\hat{\beta}} = \hat{S}^2_{ост}(X^T X)^{-1}$ (т.е. дисперсий $S^2_{\hat{\beta}_j}$), что может привести к неверным статистическим выводам о значимости коэффициентов.

1.3.1 Признаки мультиколлинерности

При исследовании модели на мультиколлинеарность к ранее рассмотренным признакам необходимо добавить следующие признаки:

Внешние (косвенные) признаки мультиколинеарности

1.Некоторые коэффициенты уравнения регрессии имеют неправильные с точки зрения экономической теории знаки или неоправданно большие по абсолютной величине значения.

Формальные признаки мультиколлинеарности:

1. Среди коэффициентов корреляционной матрицы факторных признаков есть такие, которые по величине достаточно велики (больше 0,7-0,8). Это свидетельствует о возможно достаточно тесной линейной связи.

2. Достаточно высокие значения множественных коэффициентов корреляции (детерминации) одной из объясняющей переменной (X_j) на другие: $\hat{R}^2_{x_j / x_1,...,x_{j-1},x_{j+1},...,x_k} > 0{,}7$.

1.3.2 Методы устранения мультиколлинеарности

1.3.2.1 Переход к смещенным оценкам

Пусть оценка $\hat{\beta}_{мнк}$ - наилучшая среди несмещенных оценок неизвестного значения β в классе всех несмещенных оценок $\hat{\beta}$. Повторяя выборки одного и того же объема *n* из анализируемой генеральной совокупности и подсчитывая значения $\hat{\beta}^{(l)}_{мнк}$ для каждой (*l*-й) выборки по одной и той же формуле построения наилучшей несмещенной оценки (при l =1,2,..), будем иметь множество значений $\hat{\beta}_{мнк}$, по которым можно оценить, например, плотность распределения оценки $\hat{\beta}_{мнк}$ - функцию $\hat{f}_{\hat{\beta}_{мнк}}(x)$. Поступая аналогичным образом с конкурирующей смещенной оценкой $\hat{\beta}_{см}$, получим плотность ее распределения $\hat{f}_{\hat{\beta}_{см}}(x)$.

Пусть Δ - допустимый предел погрешности в оценивании истинного значения β, т.е. если $\left|\hat{\beta} - \beta\right| \le \Delta$, то оценка $\hat{\beta}$ считается «хорошей», а при $\left|\hat{\beta} - \beta\right| > \Delta$ - «плохой».

Визуальный анализ (рисунок 1.1) приводит к выводам:

- доля «плохих» оценок $\hat{\beta}_{см}$ (а она определяется, в соответствии с вероятностным смыслом кривой плотности $\hat{f}_{\hat{\beta}_{см}}(x)$, величиной заштрихованной

площади под кривой плотности $\hat{f}_{\hat{\beta}_{см}}(x)$ вне интервала $[\beta-\Delta,\beta+\Delta]$) в несколько раз меньше доли «плохих» оценок $\hat{\beta}_{мнк}$(последняя аналогично определяется заштрихованной площадью под кривой плотности $\hat{f}_{\hat{\beta}_{мнк}}(x)$ вне интервала $[\beta-\Delta,\beta+\Delta]$);

- средний квадрат ошибок при оценивании методом $\hat{\beta}_{мнк}$ (как результат интегрирования величин $(\hat{\beta}_{мнк}-\beta)^2$ с весами, определяемыми функцией плотности $\hat{f}_{\hat{\beta}_{мнк}}(x)$, т.е. $M(\hat{\beta}_{мнк}-\beta)^2=\int_{-\infty}^{+\infty}(x-\beta)^2\hat{f}_{\hat{\beta}_{мнк}}(x)\,dx$) будет превосходить средний квадрат ошибок, получаемых при оценивании с помощью смещенной оценки (т.е. величину $M(\hat{\beta}_{см}-\beta)^2=\int_{-\infty}^{+\infty}(x-\beta)^2\hat{f}_{\hat{\beta}_{см}}(x)\,dx$)

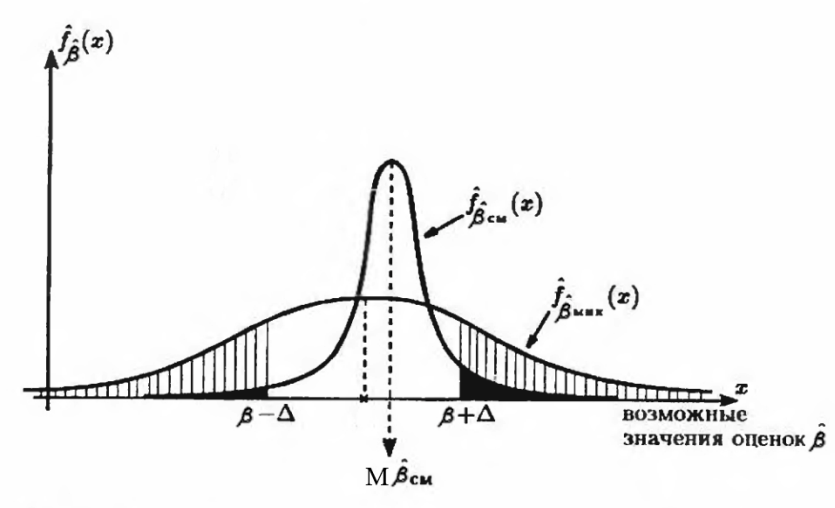


Рисунок 1.1 - Плотность распределения несмещенной ($\hat{f}_{\hat{\beta}_{мнк}}(x)$) и смещенной ($\hat{f}_{\hat{\beta}_{см}}(x)$) оценок истинного значения β неизвестного параметра

Таким образом, учитывая, что в условиях мультиколлинеарности дисперсии даже наилучших несмещенных оценок могут быть слишком большими, естественно попытаться отказаться от требования несмещенности, чтобы в более широком классе оценок найти те, которые будут обладать более высокой точностью.

Один из подходов к построению «хороших» смещенных оценок коэффициентов регрессии в условиях мультиколлинеарности называется «ридж-регрессией» («гребневой регрессией» или регуляризация) (частный случай метода регуляризации, при $\beta^0 = (0,...0)^T$ - (1.14)). Он основан на рассмотрении однопараметрического семейства несколько «подправленных» МНК-оценок, а именно оценок вида:

$$b_\tau = (X^T X + \tau \cdot E_{k+1})^{-1} X^T Y, \qquad (1.14)$$

где E_{k+1} - единичная матрица $(k+1)$ порядка;

τ - некоторое положительное число, «гребень» ($0{,}1 \le \tau \le 0{,}4$).

Добавление к диагональным элементам матрицы ($X^T X$) «гребня» τ с одной стороны, делает получаемые при этом оценки смещенными, а с другой,- превращает матрицу $X^T X$ из «плохо обусловленной» в «хорошо обусловленную». Соответственно в дальнейшем и, в частности, при вычислении средних квадратов ошибок для оценок b_τ мы не столкнемся с чрезмерно малыми значениями определителя матрицы $X^T X$ (теперь это будет уже определитель матрицы $X^T X + \tau \cdot E_{k+1}$) и связанными с этим неприятностями.

1.3.2.2 Переход к ортогональным объясняющим переменным с помощью метода главных компонент

Устранение мультиколлинеарности заключается в том, что вводят новые переменные, т.е. главные компоненты, которые являются линейными комбинациями исходных переменных, таким образом, чтобы эти новые переменные между собой оказались независимыми. Новые переменные, с одной стороны, свободны от недостатков, вызванных корреляционной зависимостью, а с другой – содержат в себе максимально возможную долю информации «старых» переменных.

Идея заключается в том, что строятся главные компоненты, рассчитывается матрицу индивидуальных («наблюденных») значений главных компонент, которая воспринимается как матрица «объект-свойство», и строится уравнение регрессии на главные компоненты. Если не удается дать содержательную интерпретацию включенным в модель главным компонентам, то осуществляется переход к исходным признакам.[1]

1.3.2.3 Метод пошаговой регрессии с включением переменных

Суть метода заключается в переходе от исходного количества объясняющих переменных $x_1, x_2, ...x_{\hat{e}}$ к меньшему числу l переменных $x^{(i_1(l))}, x^{(i_2(l))}, ..., x^{(i_l(l))}$, для которых коэффициент детерминации с результативным признаком будет максимальным.

На первом шаге $(l = 1)$ определяется первая объясняющая переменная $x^{(i_1(1))}$, которую можно назвать наиболее информативной, при условии, что в регрессионную модель Y по X мы можем включить только одну из набора объясняющих переменных. Для этого нужно оценить k моделей регрессии и в качестве наиболее информативной (наиболее существенной) объясняющей переменной выбрать ту, которой соответствует максимальный коэффициент детерминации.

На втором шаге $(l = 2)$ реализация критерия максимальности коэффициента детерминации определит уже наиболее информативную пару объясняющих переменных $x^{(i_1(1))}, x^{(i_2(2))}$, причем одна их них та, которую отобрали на предыдущем шаге. Эта пара объясняющих переменных должна будет иметь наиболее тесную статистическую связь с результативным признаком.

На третьем шаге $(l = 3)$ будет отобрана наиболее информативная тройка объясняющих переменных и т.д.

На каждом шаге рассчитываются несмещенная оценка коэффициента детерминации:

$$\widehat{R}^{*2}(l) \cong 1 - \left(1 - \widehat{R}^2(l)\right)\frac{n-1}{n-l-1}, \qquad (1.15)$$

и величина нижней доверительной границы $\widehat{R}^2_{\min}(l)$

$$\widehat{R}^2_{\min}(l) = \widehat{R}^{*2}(l) - 2\sqrt{\frac{2l(n-l-1)}{(n-1)(n^2-1)}}\left(1-\widehat{R}^2(l)\right). \qquad (1.16)$$

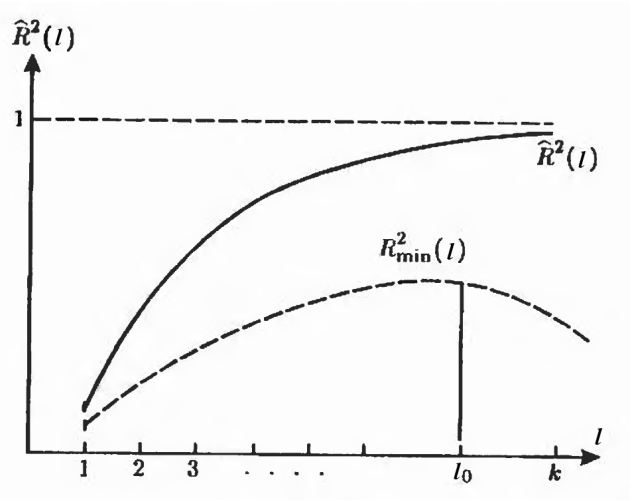


Рисунок 1.2 – Зависимость нижней доверительной границы коэффициента детерминации от числа предикторов (пунктирная кривая)

Предполагается выбирать в качестве оптимального числа l_0 объясняющих переменных регрессионной модели значение l, при котором величина $R^2_{\min}(l)$ достигает своего максимума.

1.4 Вопросы и задания, выносимые на семинарские занятия

1. Дать определение плохо обусловленной системы линейных уравнений.
2. Признаки плохой обусловленности.
3. К каким последствиям может привести плохо обусловленная систем линейных уравнений при вычислении МНК-оценок?
4. Вывести формулу для оценки коэффициентов рекуррентным методом наименьших квадратов [6, С.63-68].
5. Раскрыть понятия полной и частичной мультиколлинеарности.
6. Как выявить полную мультиколлинеарность?

7. Укажите внешние и формальные признаки мультиколлинеарности.

8. Переход от несмещенных оценок, определенных по методу наименьших квадратов, к смещенным оценкам, как метод устранения мультиколлинеарности.

9. Описать алгоритм устранения мультиколлинеарности методом ридж-регрессии.

10. В каких случаях следует применять ридж-регрессию («гребневую регрессию») для устранения последствий мультиколлинеарности?

11. Какими свойствами обладают гребневые оценки коэффициентов регрессии?

12. В чем суть метода главных компонент как средства устранения мультиколлинеарности? [1, С. 521-546].

13. Можно ли устранить мультиколлинеарность, приравняв незначимые коэффициенты в уравнении регрессии к нулю?

14. В чем суть версии “всех возможных регрессий”, как метода устранения мультиколлинеарности. [1, С. 663-664].

15. Описать алгоритм устранения мультиколлинеарности методом пошаговой регрессии с включением переменных?

16. Описать алгоритм устранения мультиколлинеарности методом пошаговой регрессии с исключением переменных [4, С. 78-79].

17. Последствие неустранения мультиколлинеарности представлено в следующем ответе:

а) МНК-оценки коэффициентов линейной модели множественной регрессии являются несмещенными и состоятельными, а их ковариационной матрицы – смещенными и несостоятельными;

б) оценка ковариационной матрицы $\Sigma_{\bar{e}}$ вектора оценок является несмещенной и несостоятельной;

в) МНК-оценки коэффициентов линейной модели множественной регрессии находятся не точно, с грубыми ошибками, поскольку искажены результаты $(X^T X)^{-1}$;

г) МНК-оценки коэффициентов линейной модели множественной регрессии являются несмещенными и несостоятельными.

18. При изучении ЛММР $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$ оценка матрицы парных коэффициентов корреляции оказалась следующей:

$$\hat{R} = \begin{array}{c} \\ y \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{c} \begin{array}{cccc} y & x_1 & x_2 & x_3 \end{array} \\ \begin{bmatrix} 1 & & & \\ 0.8 & 1 & & \\ 0.7 & 0.8 & 1 & \\ 0.6 & 0.5 & 0.2 & 1 \end{bmatrix} \end{array}$$

По полученной матрице $\hat{R}$ можно предположить о наличии мультиколлинеарности, так как:

а) $r_{y,x_1} > 0.7$, $r_{y,x_2} = 0.7$;

б) $r_{y,x_1} > 0.7$;

в) $r_{x_1,x_2} > 0.7$;

г) $r_{y,x_2} = 0.7$.

2 Практическая часть

2.1 Описание лабораторной работы

Лабораторная работа включает в себя следующие этапы:

- постановку задачи;
- ознакомление с порядком выполнения работы;
- выполнение расчетов индивидуальных задач на компьютере и анализ результатов;
- подготовку письменного отчета с выводами по работе;
- защиту лабораторной работы.

2.2 Задание к лабораторной работе

На основе показателей, характеризующих социально-экономическое развитие городов и районов Оренбургской области (Приложение А), провести исследование зависимости результативного признака от объясняющих переменных на основе линейной модели множественной регрессии:

1) построить МНК-оценки коэффициентов линейной модели множественной регрессии;

2) провести анализ построенной модели на мультиколлинеарность;

3) в случае необходимости устранить мультиколлинеарность.

2.3 Порядок выполнения работы

Изучается регрессионная зависимость ожидаемой продолжительности жизни мужчин, (y, число лет), от ряда факторов:

x_1 – общий коэффициент рождаемости (на 1000 человек);

x_2 – общий коэффициент смертности (на 1000 человек)

x_3 – уровень брачности населения (на 1000 человек);

x_4 – уровень разводимости (на 1000 человек);

x_5 – коэффициент младенческой смертности (на 1000 родившихся живыми);

x_6 – соотношение денежного дохода и прожиточного минимума, (%);

x_7 – соотношении средней оплаты труда и прожиточного минимума трудоспособного населения, (%);

x_8 – численности населения с денежными доходами ниже прожиточного минимума (в % от численности населения);

x_9 –число зарегистрированных преступлений (на 100000 человек).

Зависимость будем искать в виде:

$$\tilde{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9$$

Объектом исследования выступают города и районы Оренбургской области. Предметом исследования – взаимосвязи между ожидаемой продолжительностью жизни мужчин и указанными показателями.

Информационная база представлена данными о значениях соответствующих показателей для 48 городов и районов Оренбургской области.

Оценка параметров уравнения регрессии в ППП *Statistica* представлена на рисунке 2.1.

Regression Summary for Dependent Variable: Y (МЕТ1)
R= ,86376005 R?= ,74608142 Adjusted R?= ,68594281
F(9,38)=12,406 p<,00000 Std.Error of estimate: 1,2269

N=48	Beta	Std.Err. of Beta	B	Std.Err. of B	t(38)	p-level
Intercept			63,18456	3,826168	16,51380	0,000000
X1	0,188888	0,143370	0,16758	0,127199	1,31748	0,195565
X2	-0,081505	0,085997	-0,00769	0,008113	-0,94777	0,349237
X3	0,320317	0,110088	1,13001	0,388365	2,90966	0,006018
X4	-0,457356	0,154418	-1,12798	0,380840	-2,96181	0,005249
X5	-0,098744	0,085846	-0,07117	0,061873	-1,15024	0,257231
X6	-0,167099	0,095517	-0,03406	0,019467	-1,74941	0,088293
X7	-0,152619	0,134402	-0,00965	0,008496	-1,13554	0,263262
X8	-0,187146	0,133202	-0,04272	0,030406	-1,40498	0,168150
X9	-0,266202	0,117980	-0,00130	0,000577	-2,25633	0,029885

Рисунок 2.1 - Результаты оценивания параметров линейной модели множественной регрессии

На уровне значимости 0,05 можно принять нулевую гипотезу о том, что распределение регрессионных остатков не отличаются от нормального. (Рисунок 2.2)

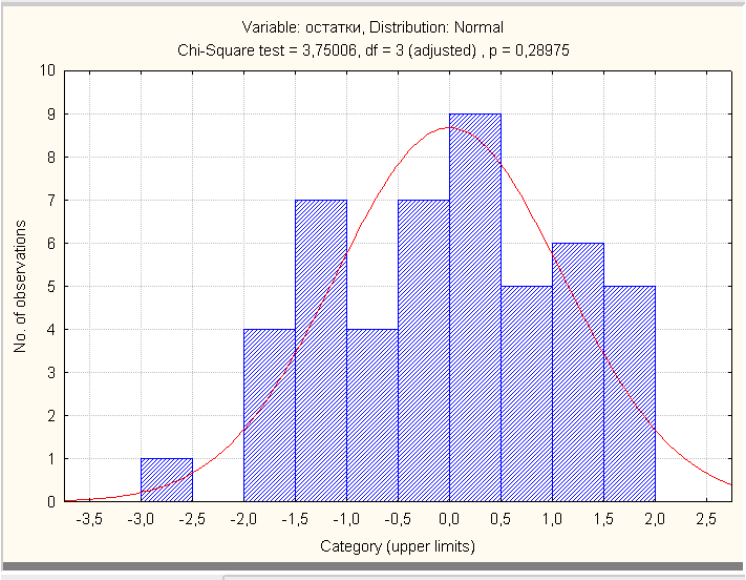


Рисунок 2.2 – График распределения регрессионных остатков

Оценка уравнения регрессии выглядит следующим образом:

$$\hat{y}=63{,}18+0{,}17x_1-0{,}008x_2+1{,}13x_3-1{,}13x_4-0{,}07x_5-0{,}034x_6-0{,}0097x_7-0{,}04x_8-0{,}001x_9$$

(0,13) (0,008) (0,39) (0,38) (0,06) (0,02) (0,008) (0,03) (0,005)

Внешние (косвенные) признаки мультиколинеарности

1. Согласно полученной модели при увеличении соотношения денежного дохода и прожиточного минимума на 1% ожидаемая продолжительность жизни мужчин уменьшится в среднем на 0,034 (коэффициент при переменной x_6 имеет отрицательный знак), что противоречит экономическому смыслу.

2. Среди коэффициентов уравнения регрессии много (коэффициенты при x_1, x_2, x_5, x_6, x_7, x_8) незначимых, а модель сама значима.

3.Среднеквадратические ошибки S_{b_j} оказались того же порядка, что и коэффициенты регрессии при переменных x1, x2, x5, x6, x7, x8. Это свидетельствует о том, что коэффициенты при соответствующих переменных могут иметь доверительный интервал, включающий в себя точку 0.

Формальные признаки мультиколлинеарности

1) Для вычисления оценки матрицы парных коэффициентов корреляции в окне множественная регрессия установим флажок в поле **Review descriptive statistics, correlations matrix.** После нажатия на кнопку **ОК** на экране откроется окно.

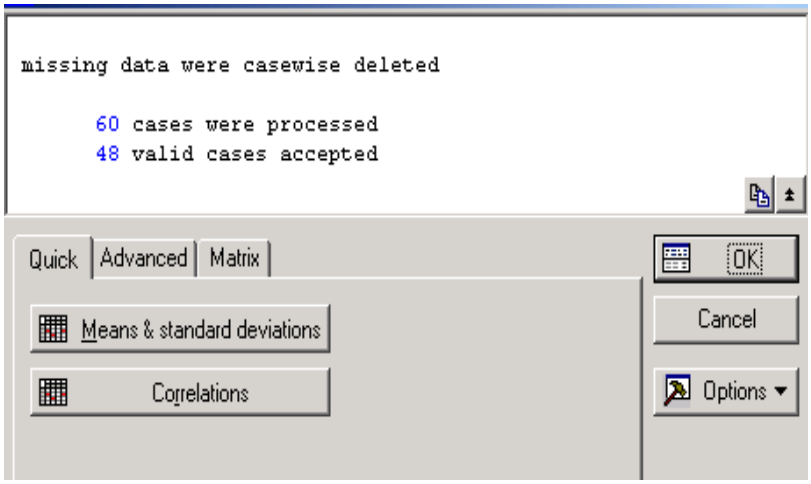


Рисунок 2.3 – Окно для вычисления оценки матрицы парных коэффициентов корреляции

В открывшемся окне нажимаем кнопку **Correlations** для вычисления оценки матрицы парных коэффициентов корреляции.

Variable	Correlations (MET1)									
	X1	X2	X3	X4	X5	X6	X7	X8	X9	Y
X1	1,000000	-0,163471	-0,033347	-0,734697	0,065213	-0,303998	-0,485993	0,566683	-0,554129	0,687535
X2	-0,163471	1,000000	-0,119662	-0,005496	-0,028424	0,130073	0,136329	-0,155531	0,107924	-0,187556
X3	-0,033347	-0,119662	1,000000	0,324062	-0,155090	0,320078	0,214232	-0,113756	-0,310372	0,208605
X4	-0,734697	-0,005496	0,324062	1,000000	-0,155051	0,338975	0,513321	-0,463331	0,481303	-0,652968
X5	0,065213	-0,028424	-0,155090	-0,155051	1,000000	-0,163029	0,025993	0,054247	0,031511	-0,058139
X6	-0,303998	0,130073	0,320078	0,338975	-0,163029	1,000000	0,388647	-0,368066	0,101189	-0,288899
X7	-0,485993	0,136329	0,214232	0,513321	0,025993	0,388647	1,000000	-0,745229	0,148294	-0,389195
X8	0,566683	-0,155531	-0,113756	-0,463331	0,054247	-0,368066	-0,745229	1,000000	-0,235037	0,340489
X9	-0,554129	0,107924	-0,310372	0,481303	0,031511	0,101189	0,148294	-0,235037	1,000000	-0,697877
Y	0,687535	-0,187556	0,208605	-0,652968	-0,058139	-0,288899	-0,389195	0,340489	-0,697877	1,000000

Рисунок 2.4 – Оценка матрицы парных коэффициентов корреляции

На основе вычисленной матрицы есть основания подозревать тесную связь между x_1 и x_4 ($r(x^{(1)}, x^{(4)}) = 0{,}73$) и x_7 и x_8 ($r(x^{(7)}, x^{(8)}) = -0{,}75$).

2. Более внимательное изучение этого вопроса достигается с помощью расчета значений коэффициентов детерминации $\hat{R}^2_{x^{(j)} \cdot X(j)}$ каждой из объясняющих переменных $x^{(j)}$ по всем остальным переменным $X(j) = \left(x^{(1)}, ..., x^{(j-1)}, x^{(j+1)}, ...x^{(p)}\right)$.

Для определения коэффициентов детерминации следует воспользоваться модулем множественная регрессия, где в качестве зависимой переменной выбрать $x^{(j)}$, все остальные объясняющие переменные в качестве независимых (рисунок 2.5).

Multiple Regression Results

Dependent: X1 Multiple R = ,82153305 F = 10,12115
R?= ,67491655 df = 8,39
No. of cases: 48 adjusted R?= ,60823277 p = ,000000
Standard error of estimate: 1,544474752
Intercept: 12,830009051 Std.Error: 4,356569 t(39) = 2,9450 p = ,0054

Рисунок 2.5– Оценка коэффициента детерминации переменной x_1

Все расчеты остальных коэффициенты детерминации производятся аналогичным образом. В результате получили:

$$\hat{R}^2_{x_1 / x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9} = 0.67498$$

$$\hat{R}^2_{x_2 / x_1 x_3 x_4 x_5 x_6 x_7 x_8 x_9} = 0.0965$$

$$\hat{R}^2_{x_3/x_1x_2x_4x_5x_6x_7x_8x_9} = 0.4486$$

$$\hat{R}^2_{x_4/x_1x_2x_3x_5x_6x_7x_8x_9} = 0.7198$$

$$\hat{R}^2_{x_5/x_1x_2x_3x_4x_6x_7x_8x_9} = 0.0933$$

$$\hat{R}^2_{x_6/x_1x_2x_3x_4x_5x_7x_8x_9} = 0.2676$$

$$\hat{R}^2_{x_7/x_1x_2x_3x_4x_5x_6x_8x_9} = 0.6301$$

$$\hat{R}^2_{x_8/x_1x_2x_3x_4x_5x_6x_7x_9} = 0.6234$$

Рисунок 2.6 – Результаты вычислений оценок коэффициента детерминации

Анализ оценок коэффициентов детерминации показал наличие тесной линейной связи между объясняющей переменной x_4 и всеми остальными признаками, то же самое можно сказать о переменных x_7, x_8, x_1.

3. Достаточным условием плохой обусловленности матрицы (наличия мультиколинеарности) является большое значение числа обусловленности. Для вычисления собственных чисел матрицы X^TX воспользуемся функциональными возможностями программы Mathcad.

Сначала матрицу X формируем в Excel, сохраняем в текстовом формате (с расширением *.txt), затем открываем Mathcad, в меню **Insert** выбираем пункт **Components** (рисунок 2.7).

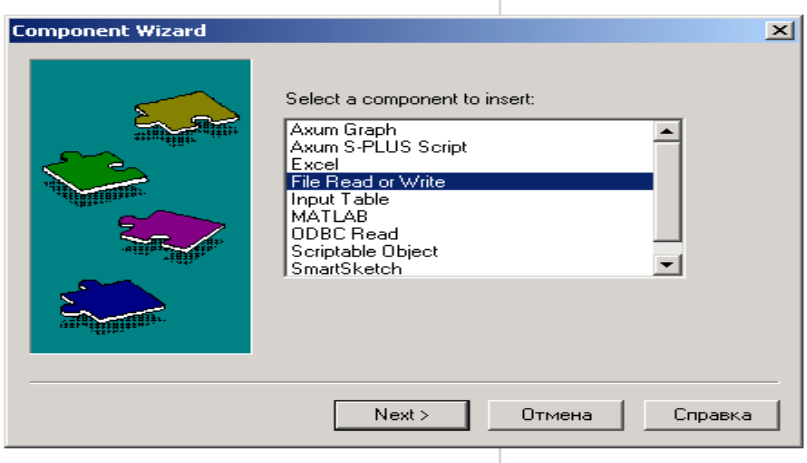


Рисунок 2.7 – Вид окна Component Wizard

В появившемся окне находим пункт **File Read or Write** и нажимаем на кнопку **Next – Далее**. На экране открывается следующее окно (рисунок 2.8).

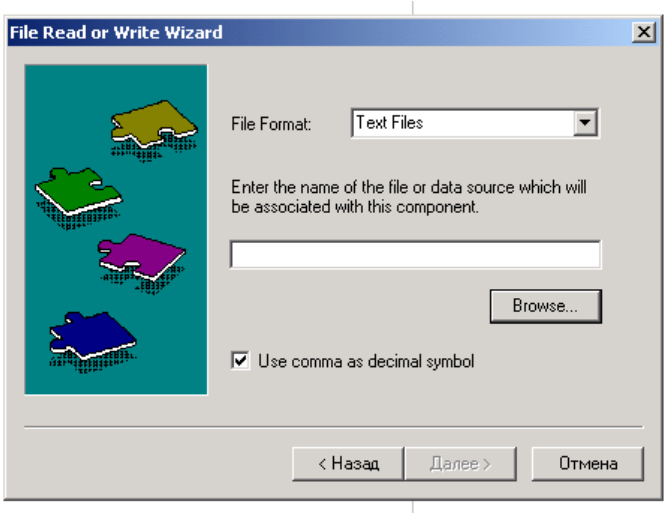


Рисунок 2.8– Вид окна File Read or Write Wizard

В окне **File Read or Write Wizard** нажимаем на кнопку **Browse - Обзор** и открываем текстовый файл, в котором сохранили матрицу X. Выбрав нужный файл, нажимаем на кнопку **Готово** (рисунок 2.9).

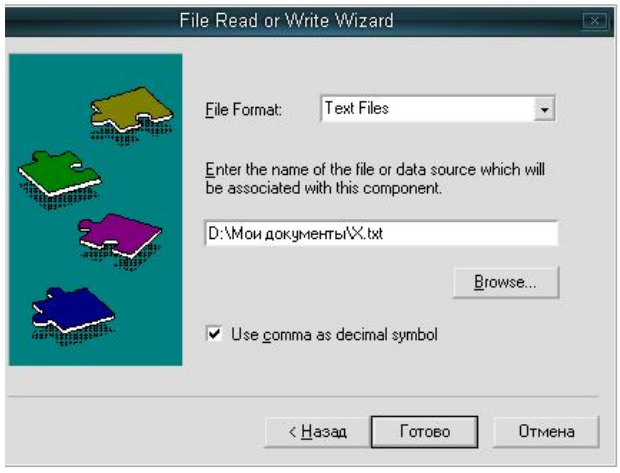


Рисунок 2.9 - Вид окна File Read or Write Wizard

В появившемся окне, полученной матрицы присваиваем имя (рисунок 2.10).

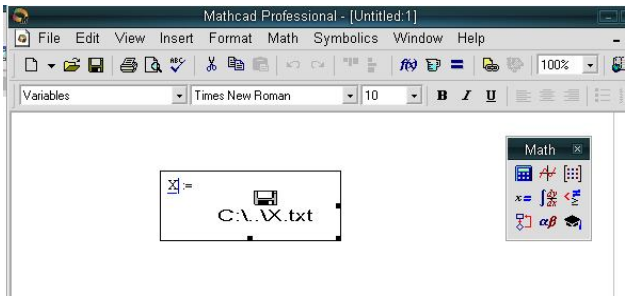
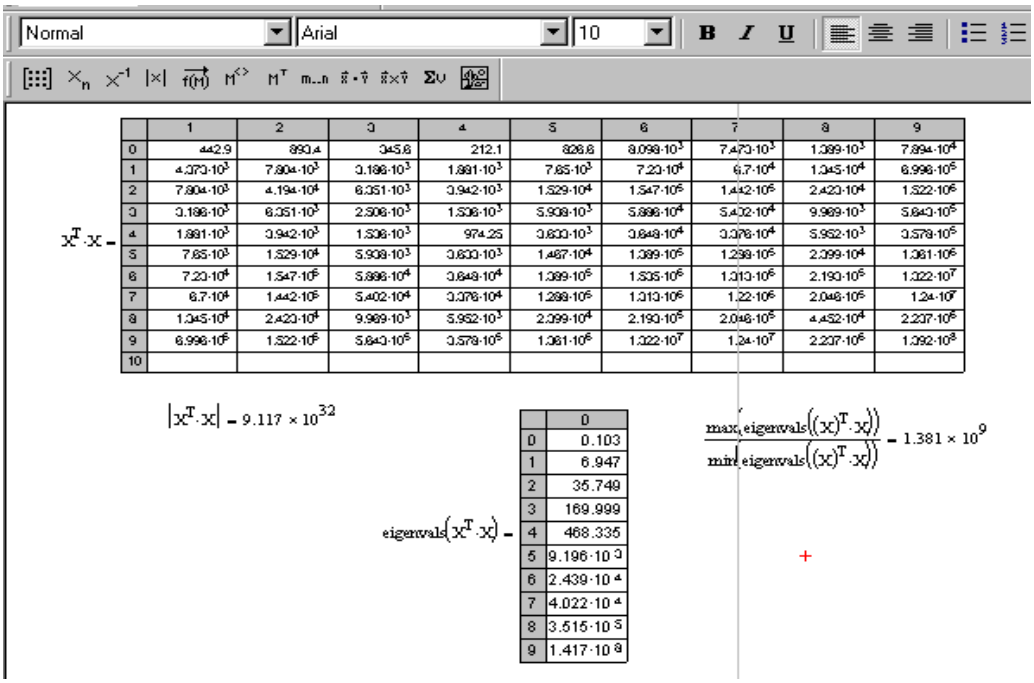


Рисунок 2.10

Вычислим собственные числа матрицы X^TX в программе Mathcad, воспользовавшись функцией eiganvals (рисунок 2.11).



	1	2	3	4	5	6	7	8	9
0	442.9	893.4	345.6	212.1	828.6	$8.098 \cdot 10^3$	$7.473 \cdot 10^3$	$1.389 \cdot 10^3$	$7.894 \cdot 10^4$
1	$4.373 \cdot 10^3$	$7.804 \cdot 10^3$	$3.186 \cdot 10^3$	$1.891 \cdot 10^3$	$7.65 \cdot 10^3$	$7.23 \cdot 10^4$	$6.7 \cdot 10^4$	$1.345 \cdot 10^4$	$6.996 \cdot 10^5$
2	$7.804 \cdot 10^3$	$4.194 \cdot 10^4$	$6.351 \cdot 10^3$	$3.942 \cdot 10^3$	$1.529 \cdot 10^4$	$1.547 \cdot 10^5$	$1.442 \cdot 10^5$	$2.423 \cdot 10^4$	$1.322 \cdot 10^6$
3	$3.186 \cdot 10^3$	$6.351 \cdot 10^3$	$2.906 \cdot 10^3$	$1.536 \cdot 10^3$	$5.908 \cdot 10^3$	$5.898 \cdot 10^4$	$5.402 \cdot 10^4$	$9.969 \cdot 10^3$	$5.643 \cdot 10^5$
4	$1.891 \cdot 10^3$	$3.942 \cdot 10^3$	$1.536 \cdot 10^3$	974.25	$3.633 \cdot 10^3$	$3.648 \cdot 10^4$	$3.376 \cdot 10^4$	$5.952 \cdot 10^3$	$3.578 \cdot 10^5$
5	$7.65 \cdot 10^3$	$1.529 \cdot 10^4$	$5.908 \cdot 10^3$	$3.633 \cdot 10^3$	$1.467 \cdot 10^4$	$1.389 \cdot 10^5$	$1.298 \cdot 10^5$	$2.299 \cdot 10^4$	$1.361 \cdot 10^6$
6	$7.23 \cdot 10^4$	$1.547 \cdot 10^5$	$5.898 \cdot 10^4$	$3.648 \cdot 10^4$	$1.389 \cdot 10^5$	$1.335 \cdot 10^6$	$1.213 \cdot 10^6$	$2.193 \cdot 10^5$	$1.322 \cdot 10^7$
7	$6.7 \cdot 10^4$	$1.442 \cdot 10^5$	$5.402 \cdot 10^4$	$3.376 \cdot 10^4$	$1.298 \cdot 10^5$	$1.213 \cdot 10^6$	$1.22 \cdot 10^6$	$2.048 \cdot 10^5$	$1.24 \cdot 10^7$
8	$1.345 \cdot 10^4$	$2.423 \cdot 10^4$	$9.969 \cdot 10^3$	$5.952 \cdot 10^3$	$2.299 \cdot 10^4$	$2.193 \cdot 10^5$	$2.048 \cdot 10^5$	$4.452 \cdot 10^4$	$2.237 \cdot 10^6$
9	$6.996 \cdot 10^5$	$1.322 \cdot 10^6$	$5.643 \cdot 10^5$	$3.578 \cdot 10^5$	$1.361 \cdot 10^6$	$1.322 \cdot 10^7$	$1.24 \cdot 10^7$	$2.237 \cdot 10^6$	$1.392 \cdot 10^8$
10									

	0
0	0.103
1	6.947
2	35.749
3	169.999
4	468.335
5	$9.196 \cdot 10^3$
6	$2.439 \cdot 10^4$
7	$4.022 \cdot 10^4$
8	$3.515 \cdot 10^5$
9	$1.417 \cdot 10^8$

Рисунок 2.11– Результаты вычислений в программе Mathcad

Таким образом, можно говорить о наличии мультиколлинеарности объясняющих переменных $x_1,\ldots,x_9$.

В случае, если между объясняющими переменными существует частичная мультиколлинеарность, то оценки коэффициентов линейной модели, полученные по МНК, становятся неустойчивыми, незначительное изменение состава выборки или состава объясняющих переменных может вызвать кардинальное изменение модели, что делает модель непригодной для практических целей. Наиболее распространенные в таких случаях приемы оценивания параметров регрессионной модели: методы пошаговой регрессии, использование гребневой регрессии (ридж-регрессии), переход от первоначальных переменных к их главным компонентам [1,2]. Все вышеприведенные методы реализуются в ППП Statistica. Рассмотрим некоторые методы устранения регрессии, используя модуль «множественная регрессия».

Установка флажка в поле **Advanced options** модуля множественная регрессия позволит перейти к диалоговому окну **Model Defenition,** открывающему возможность выбора метода анализа, среди которых методы пошаговой регрессии и гребневой. В прокручиваемом списке методов можно выбрать один из методов пошаговой регрессии.

Методы пошаговой регрессии позволяют из множества независимых переменных отобрать только те, которые наиболее значимы для адекватного описания многопараметрической регрессии. В модуле реализованы две процедуры отбора переменных, каждая из которых может давать различный конечный набор переменных: последовательное включение **(Forward stepwise)** и последовательное исключение **(Backward stepwise).** Гребневая регрессия используется для получения более устойчивых оценок параметров регрессионной модели в условиях мультиколлинеарности переменных.

Устраним мультиколлинеарность методом пошагового включения:

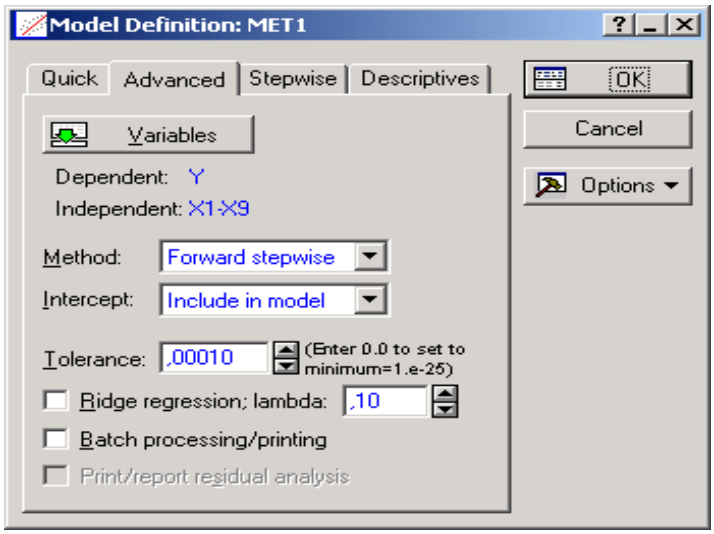


Рисунок 2.12 – Выбор метода оценивания параметров регрессионной модели

Результаты расчетов приведены в виде отчета на рисунке 2.13.

Regression Summary for Dependent Variable: Y (MET1)
R= ,82174263 R?= ,67526096 Adjusted R?= ,65311966
F(3,44)=30,498 p<,00000 Std.Error of estimate: 1,2894

N=48	Beta	Std.Err. of Beta	B	Std.Err. of B	t(44)	p-level
Intercept			60,41614	2,771840	21,79640	0,000000
X9	-0,313403	0,118526	-0,00153	0,000579	-2,64416	0,011308
X4	-0,601357	0,119100	-1,48313	0,293737	-5,04917	0,000008
X3	0,306210	0,109821	1,08024	0,387426	2,78826	0,007798

Рисунок 2.13– Результаты оценивания параметров линейной модели множественной регрессии методом пошаговой регрессии с включением переменных

Были исследованы также регрессионные остатки, анализ которых показал нормальность их распределения (рисунок 2.14).

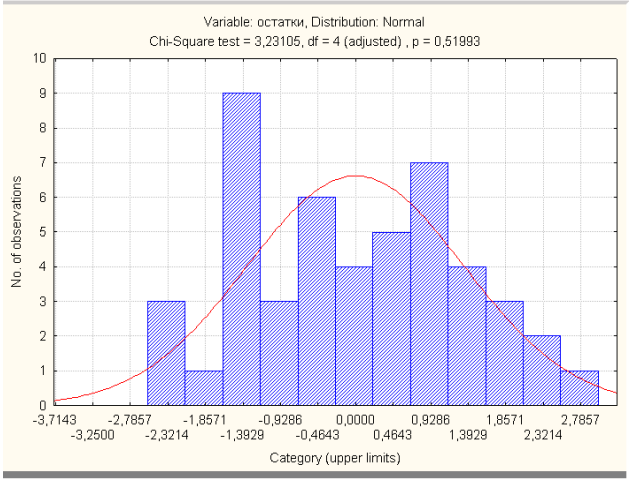


Рисунок 2.14– Гистограмма распределения регрессионных остатков

В результате проведения пошаговой регрессии получили следующую оценку уравнения регрессии:

$$\hat{y} = \underset{(2,77)}{60,42} + \underset{(0,39)}{1,88}x_3 - \underset{(0,29)}{1,48}x_4 - \underset{(0,0006)}{0,88}x_9, \qquad \widehat{R}^2 = 0,675,\ \hat{S}^2_{ост} = 1,66$$

Оценка уравнения регрессии значима т.к. нулевая гипотеза отклонена; коэффициенты при переменных также значимы. Коэффициент детерминации составил 0,675, т.е. 67,5% доли вариации результирующей переменной объясняется переменными x_3, x_4 и x_9, а 32,5% доли вариации, вероятно, объясняется неучтенными в модели факторами.

Согласно полученной модели, можно сделать вывод о том, что увеличение количества браков приводит к росту ожидаемой продолжительности жизни мужчин в среднем на 1,08 лет, при росте количества разводов ожидаемая продолжительность жизни мужчин в среднем сокращается на 1,48 лет, при увеличении числа зарегистрированных преступлений ожидаемая продолжительность жизни мужчин в среднем также сокращается на 0,0015 лет.

Устраним мультиколлинеарность методом пошагового исключения:

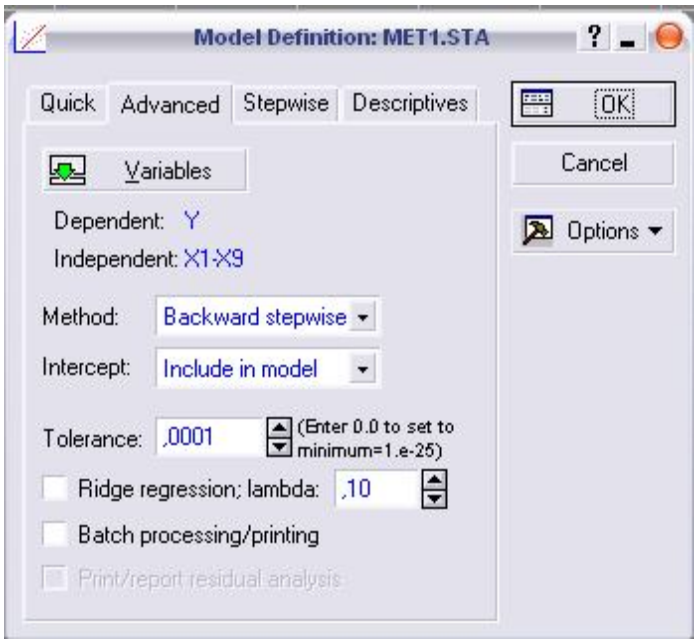


Рисунок 2.15 – Выбор метода оценивания параметров регрессионной модели

Результаты расчетов приведены в виде отчета на рисунке 2.16.

Regression Summary for Dependent Variable: Y (MET1.STA)
R= ,78972141 R?= ,62365991 Adjusted R?= ,60693368
F(2,45)=37,286 p<,00000 Std.Error of estimate: 1,3725

N=48	Beta	Std.Err. of Beta	B	Std.Err. of B	t(45)	p-level
Intercept			55,96935	2,345478	23,86266	0,000000
X3	0,469513	0,096667	1,65634	0,341019	4,85703	0,000015
X4	-0,805119	0,096667	-1,98567	0,238409	-8,32883	0,000000

Рисунок 2.16– Результаты оценивания параметров линейной модели множественной регрессии методом пошаговой регрессии с исключением переменных

Были исследованы также регрессионные остатки, анализ которых показал нормальность их распределения (рисунок 2.17).

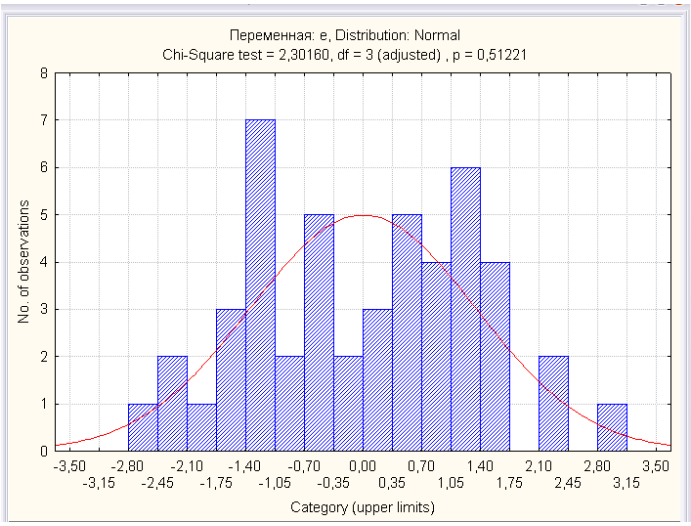


Рисунок 2.17– Гистограмма распределения регрессионных остатков

В результате проведения пошаговой регрессии получили следующую оценку уравнения регрессии:

$$\hat{y} = \underset{(2,35)}{55,97} + \underset{(0,34)}{1,66}\, x_3 - \underset{(0,24)}{1,99}\, x_4 , \qquad \widehat{R}^2 = 0,624 , \quad \hat{S}^2_{ост} = 1,884$$

Оценка уравнения регрессии значима т.к. нулевая гипотеза отклонена; коэффициенты при переменных также значимы. Коэффициент детерминации составил 0,624, т.е. 62,4% доли вариации результирующей переменной объясняется переменными x_3, x_4 , а 37,6% доли вариации, вероятно, объясняется неучтенными в модели факторами.

Согласно полученной модели, можно сделать вывод о том, что увеличение количества браков приводит к росту ожидаемой продолжительности жизни мужчин в среднем на 1,66 лет, при росте количества разводов ожидаемая продолжительность жизни мужчин в среднем сокращается на 1,99 лет.

Таким образом, получены две модели ожидаемой продолжительности жизни мужчин, из которых нужно по экономическим и статистическим соображениям выбрать наилучшую. По статистическим критериям наиболее адекватна первая модель: ей соответствует минимальное значение остаточной дисперсии и наибольшее значение коэффициента детерминации. К тому в первой модели,

наряду с общими для моделей факторами x_3, x_4, присутствует переменная x_9, которая оказывает влияние на результативный признак.

Таким образом, после реализации алгоритма пошагового регрессионного анализа выбираем окончательное уравнение регрессии:

$$\hat{y} = \underset{(2,77)}{60,42} + \underset{(0,39)}{1,88}x_3 - \underset{(0,29)}{1,48}x_4 - \underset{(0,0006)}{0,88}x_9, \qquad \widehat{R}^2 = 0,675,\ \hat{S}^2_{ост} = 1,66$$

Для реализации метода гребневой регрессии (ридж-регрессии), необходимо в окне **Model Defenition** (рисунок 2.12) установить флажок в поле **Ridge regression** и указать величину «гребня», «хребта» в диапазоне значений от 0,1 до 0,4 [1].

2.4 Содержание письменного отчета

Отчет должен быть оформлен на листах формата А4 с титульным листом, оформленным соответствующим образом и содержать следующее:

1) постановку задачи с вариантом выборок;

2) краткое изложение теории по исследованию ЛММР на мультиколлинеарность;

3) результаты компьютерной обработки данных;

4) анализ полученных результатов;

5) выводы по полученным результатам.

2.5 Вопросы к защите лабораторной работы

1) Сформулируйте постановку задачи лабораторной работы
2) Запишите результаты наблюдений в виде матрицы Y и матрицы X.
3) Какое программное обеспечение использовалось для решения задачи?
4) Какие внешние признаки мультиколлинеарности позволили заподозрить ее наличие?

5) Какие формальные признаки позволили заподозрить наличие мультиколлинеарности?

6) При анализе линейной модели регрессии на мультиколлинеарность в матрице парных коэффициентов корреляции между объясняющими переменными не оказалось элементов, превышающих 0,7 по модулю. Можно ли в этом случае говорить об отсутствии мультиколлинеарности?

7) Каким методом была устранена мультиколлинеарность в лабораторной работе?

8) Чему равен коэффициент детерминации? Что он характеризует?

Список использованных источников

1 Айвазян, С.А. Прикладная статистика и основы эконометрики: учебник для вузов/ С.А. Айвазян, В.С. Мхитарян. – М.: ЮНИТИ, 1998. – 1022 с.

2 Большаков, А.А. Методы обработки многомерных данных и временных рядов: учебное пособие для вузов / А.А. Большаков, Р.Н. Каримов – М.: Горячая линия – Телеком, 2007. – 522 с.

3 Магнус, Я.Р. Эконометрика. Начальный курс: учебник/ Я.Р. Магнус, П.К. Катышев, А.А. Пересецкий. – 6-е изд., перераб. и доп. – М.: Дело, 2004. – 57 с.

4 Мхитарян, В.С. Эконометрика: учебник / под ред. В.С. Мхитаряна. – М: Проспект, 2009.-384 с.

5 Тихомиров, Н.П. Эконометрика: учебник/ Н.П. Тихомиров, Е.Ю. Дорохина. – М.: Издательство «Экзамен», 2003. – 512 с.

6 Чураков, Е.П. Математические методы обработки экспериментальных данных в экономике: учеб.пособие/ Е.П. Чураков – М.: Финансы и статистика, 2004. – 240с.:ил.

Приложение А
(обязательное)

Исходные данные для анализа

Таблица А.1 - Значения социально-экономических показателей, характеризующих города и районы Оренбургской области

Номер объекта	Муниципальные образования	X1	X2	X3	X4	X5	X6	X7	X8
1	2	3	4	5	6	7	8	9	10
1	Абдулинский	716,59	0	50	71,8	0,00	21,90	53,09	16,64
2	Адамовский	4791,44	10296,08	16,7	46,2	0,00	8,42	61,90	22,39
3	Акбулакский	5677,90	1478,16	52,9	47,3	0,18	11,80	62,56	21,02
4	Александровский	1571,20	377,53	45,5	0	0,00	14,55	60,55	20,02
5	Асекеевский	3704,46	642,03	40,9	57,5	0,54	13,58	58,68	18,11
6	Беляевский	3304,59	341,69	72,7	22,9	0,22	17,10	61,23	19,72
7	Бугурусланский	4367,39	261,84	58,3	28,7	0,00	14,97	59,99	17,51
8	Бузулукский	2127,96	1111,62	52,6	72,8	0,36	10,76	58,65	17,74
9	Гайский	13657,15	0,00	16,7	15,1	0,00	15,65	59,70	20,32
10	Грачевский	2252,99	1385,21	55,6	73	0,13	10,93	60,46	18,07
11	Домбаровский	2242,38	508,48	81,8	25,6	0,00	6,73	62,41	22,86
12	Илекский	2803,27	505,37	42,9	4,9	3,53	11,41	59,78	19,15
13	Кваркенский	1984,05	3094,73	41,2	35,6	1,49	10,29	60,23	21,36
14	Красногвардейский	3618,35	1314,39	25	8	0,44	11,35	60,34	20,93
15	Кувандыкский	2438,19	0,00	52,9	33,3	0,00	14,47	59,02	20,38
16	Курманаевский	2074,29	0,00	81,2	57,9	1,58	21,11	60,02	17,32
17	Матвеевский	2172,78	102,99	11,1	8,3	0,00	14,66	58,83	18,09
18	Новоорский	10893,40	82540,63	61,9	8,4	0,00	12,86	61,61	19,81
19	Новосергиевский	5723,31	4935,74	40,5	20,6	0,16	11,94	59,04	19,02
20	Октябрьский	4967,20	444,28	31,2	46,7	0,13	12,81	60,50	18,03
21	Оренбургский	20071,10	25359,07	31,1	5,8	0,01	4,83	63,93	18,41
22	Первомайский	1795,32	3312,16	17,6	18,4	1,09	18,87	62,33	22,06
23	Переволоцкий	3561,15	86,88	30,8	0,8	0,32	14,63	60,62	18,92
24	Пономаревский	2217,02	184,32	81,8	52,2	0,00	16,39	57,82	16,93
25	Сакмарский	4551,40	374,80	14,3	1,4	0,03	9,11	63,10	18,20
26	Саракташский	3384,80	3525,57	40	13,6	0,00	2,30	59,67	18,21
27	Светлинский	3775,83	12159,95	64,3	12,6	0,43	10,12	61,36	21,16
28	Северный	2264,20	0,00	21,4	58,8	0,00	7,12	58,67	17,78
29	Соль-Илецкий	1047,46	358,37	81	43,1	0,88	8,19	59,74	22,86
30	Сорочинский	2833,94	13,55	33,3	40,9	0,00	26,12	56,48	20,25
31	Ташлинский	6881,07	5509,45	55,6	12,7	1,08	8,45	61,57	20,67
32	Тоцкий	1755,21	159,03	54,2	13	0,04	16,18	72,49	13,84
33	Тюльганский	3196,66	1403,25	50	14,2	2,48	15,21	62,25	18,85
34	Шарлыкский	3649,02	299,66	26,3	43,6	0,05	12,45	57,57	16,84
35	Ясненский	7148,83	0,00	100	84,4	0,63	34,94	60,79	24,15
36	Абдулино	2784,39	3277,23	25	27,4	0,02	7,49	64,93	19,16
37	Бугуруслан	4229,97	191924,71	38,9	20,6	0,05	15,35	62,18	18,16
38	Бузулук	61679,53	240951,01	27,7	0,3	0,09	6,19	64,27	15,81
39	Гай	27338,48	106449,61	10	7,8	0,04	1,82	65,93	15,24
40	Кувандык	2012,36	20786,78	27,3	0,7	0,00	7,73	63,44	17,12
41	Медногорск	11170,01	27319,93	31,2	1,5	0,00	18,05	63,00	16,65

Продолжение таблице А.1

1	2	3	4	5	6	7	8	9	10
42	Новотроицк	29743,64	217430,62	39,5	14,8	0,00	13,14	60,46	15,03
43	Оренбург	21460,65	8736,67	22,3	6,7	0,01	25,37	64,76	15,29
44	Орск	4301,33	139154,85	28,8	14,1	0,00	2,96	66,55	15,09
45	Соль-Илецк	4401,00	12593,97	42,9	6,4	0,00	0,00	63,42	16,20
46	Сорочинск	3446,14	315863,20	12,5	3,3	0,02	14,08	63,03	20,16
47	Ясный	3539,32	29399,98	50	46,3	0,00	7,82	63,26	18,17

Таблица А.2 - Значения социально-экономических показателей, характеризующих города и районы Оренбургской области

Номер объекта	Муниципальные образования	X9	X10	X11	X12	X13	X14	X15	X16
1	2	3	4	5	6	7	8	9	10
1	Абдулинский	751	0	31,35	2226,11	0,00	5158,03	329,71	99,63
2	Адамовский	2910	355	2,25	135701,34	-0,02	5908,71	2008,04	144,78
3	Акбулакский	1357	263	0,11	-8567,61	0,61	4379,21	1458,21	142,39
4	Александровский	969	26	0,17	-36522,68	0,00	6962,00	1821,81	140,83
5	Асекеевский	1643	141	0,37	17280,55	3,50	4529,49	2005,23	124,18
6	Беляевский	1502	173	0,17	-23702,21	0,00	5330,26	1583,71	131,50
7	Бугурусланский	2158	43	0,07	2327,56	-3,36	6830,25	1283,09	150,45
8	Бузулукский	1829	574	0,77	-20227,66	2,15	3813,11	1556,33	189,27
9	Гайский	1622	67	0,15	90494,93	0,00	5260,96	1543,57	132,57
10	Грачевский	1306	90	0,22	-21387,79	-4,28	5562,88	2376,39	179,24
11	Домбаровский	716	21	0,60	-66252,50	-13,30	4790,24	1855,85	158,35
12	Илекский	2098	151	2,94	-38968,18	-7,22	4117,78	1780,37	113,90
13	Кваркенский	1904	319	1,83	95392,98	9,08	4916,49	1746,81	137,01
14	Красногвардейский	814	215	0,92	-6880,28	-20,36	5483,97	1738,92	169,90
15	Кувандыкский	1529	50	0,73	-12601,50	0,00	2805,11	660,66	102,55
16	Курманаевский	1223	40	0,08	-19203,87	0,00	5175,18	1614,86	183,10
17	Матвеевский	1223	51	0,12	27154,79	0,00	8012,55	1479,75	114,97
18	Новоорский	531	1468	1,11	-88359,71	-20,64	6883,78	2791,04	219,82
19	Новосергиевский	2747	998	1,52	53771,06	57,54	12916,69	2447,89	164,51
20	Октябрьский	2019	221	0,07	6046,84	0,00	7530,85	2109,25	166,12
21	Оренбургский	2965	1984	6,90	222587,21	40,10	10051,88	9987,58	414,60
22	Первомайский	1023	161	0,11	11834,97	2,66	4820,43	1600,04	191,82
23	Переволоцкий	1548	110	2,12	5089,93	0,00	7200,79	2466,00	149,14
24	Пономаревский	609	27	17,11	-4358,73	0,00	9429,51	1663,79	153,87
25	Сакмарский	1415	714	2,07	47042,61	-18,18	5460,64	2252,94	171,96
26	Саракташский	2855	554	0,39	18636,05	-17,28	7163,74	2442,46	148,52
27	Светлинский	1261	739	0,28	-31576,96	1,10	5828,43	3104,97	173,23
28	Северный	842	55	0,29	12573,25	0,00	10527,40	2041,39	164,51
29	Соль-Илецкий	2160	238	1,42	-27755,42	2,87	9624,39	749,79	87,13
30	Сорочинский	2366	611	0,23	-41927,34	0,00	3258,68	992,49	125,95
31	Ташлинский	3706	448	0,43	19211,06	5,86	8003,63	1932,97	113,73

Продолжение таблицы А.2

1	2	3	4	5	6	7	8	9	10
32	Тоцкий	913	163	0,17	2703,44	-13,48	5710,41	1623,23	155,14
33	Тюльганский	1395	260	0,14	-1805,00	-5,70	6922,20	2492,35	135,65
34	Шарлыкский	1386	122	2,09	29131,25	0,00	8597,17	1848,44	140,58
35	Ясненский	565	0	3,86	-70126,16	0,00	5709,72	2688,39	108,26
36	Абдулино	0	1062	10,65	28860,67	3,23	20528,50	5245,52	253,70
37	Бугуруслан	0	3504	0,76	93608,09	3,43	12009,56	6175,02	254,24
38	Бузулук	0	12002	6,22	8735783,87	41,62	16533,55	8125,77	362,54
39	Гай	0	8180	0,38	1526864,98	4,32	10133,53	7199,35	316,32
40	Кувандык	0	2192	0,35	27775,53	1,38	11328,86	7323,76	192,27
41	Медногорск	0	4033	2,18	321986,13	6,97	10895,88	5295,27	248,44
42	Новотроицк	199	24413	13,08	1476312,51	18,23	13505,43	6902,37	306,62
43	Оренбург	840	52066	26,49	144294,97	8,97	57813,57	10336,03	284,29
44	Орск	0	24492	0,47	710226,74	2,28	15867,96	6455,67	339,72
45	Соль-Илецк	0	1035	0,15	29769,64	6,72	14853,78	7020,26	221,38
46	Сорочинск	0	879	0,85	40383,38	23,39	14838,68	7599,74	241,49
47	Ясный	0	4084	0,87	282356,73	34,75	9427,45	7508,34	263,32

Таблица А.3 – Наименование показателей

Обозначения	Наименование показателя
X1	Объем инвестиций в основной капитал на душу населения, рублей
X2	Объем промышленной продукции на душу населения, рублей
X3	Удельный вес убыточных предприятий и организаций, в процентах от общего числа предприятий
X4	Просроченная кредиторская задолженность предприятий, в процентах от общей задолженности
X5	Задолженность организаций по заработной плате, в процентах от общего фонда заработной платы
X6	Уровень безработицы, в процентах от населения в трудоспособном возрасте
X7	Доля населения в трудоспособном возрасте в общей численности населения, в процентах
X8	Доля лиц моложе трудоспособного возраста, в общей численности населения, в процентах
X9	Среднегодовая численность работников, занятых в сельскохозяйственном производстве, человек
X10	Среднегодовая численность работников, занятых в промышленности, человек
X11	Число зарегистрированных иностранных рабочих, в промилле от численности населения в трудоспособном возрасте
X12	Сальдированный финансовый результат (прибыль минус убыток) на одно предприятие, рублей
X13	Уровень рентабельности реализованной продукции сельского хозяйства в сельскохозяйственных организациях, в процентах
X14	Оборот розничной торговли на душу населения, рублей
X15	Объем платных услуг на душу населения, рублей
X16	Соотношение среднемесячной номинальной начисленной заработной платы работников с величиной прожиточного минимум, в процентах

Таблица А.4 – Варианты заданий

Номер варианта	Результативный признак, (обозначить Y)	Номера факторных признаков, X
1	X1	4,6,10,11,14
2	X1	5,10,11,14,15
3	X1	2,10,11,13,14
4	X1	6,7,10,12,15
5	X1	4,5,6,10,15
6	X1	3,10,11,12,15
7	X1	2,12,13,14,15
8	X1	2,9,11,14,15
9	X1	3,510,12,13
10	X1	4,5,14,15,16
11	X2	3,12,13,14,15
12	X2	4,7,11,12,13
13	X2	4,10,12,14,16
14	X2	1,9,13,15,16
15	X2	9,10,12,14,16
16	X2	9,10,13,15,16
17	X2	1,4,6,7,15
18	X3	1,4,6,8,13
19	X4	3,6,7,15,16
20	X4	2,3,6,15,16