МИНИСТЕРСВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ 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)

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

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

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА (БАКАЛАВРСКАЯ РАБОТА)

на тему Корреляционно-регрессионный анализ показателей сельскохозяйственного производства

Студент К.Д. Джоракулыев
(И.О. Фамилия) (личная подпись)

Руководитель доцент, Н.А. Сосина
(ученая степень, звание, И.О. Фамилия)

Тольятти 2020

Аннотация

Тема бакалаврской работы: «Корреляционно-регрессионный анализ показателей сельскохозяйственного производства»

В данной выпускной квалификационной работе построена многофакторная модель сельскохозяйственного предприятия на примере предприятия «Turkmenistan» по выращивание хлопка в Лебапском велаяте Туркменистана. Приводятся оценки качества регрессионной модели, осуществляется количественный анализ основных показателей сельскохозяйственного предприятия, выполнена проверка модели на адекватность, разработана программа для построение многофакторный регрессионной модели.

В работе выполнены следующие задачи:

- Построена многофакторная модель для результативного признака на примере конкретного сельскохозяйственного предприятия, выполнена проверка модели на адекватность;
- Разработано программное обеспечение для построения многофакторной модели.

Для реализации в качестве языка программирования использовался Python с использованием библиотек: numpy, statsmodels, math. Проведено исследование с использованием данной программы и получены остатки отображающие качество прогноза предложенной модели.

Данная бакалаврская работа состоит из пояснительной записки на 45 стр., включая 14 рисунков.

Abstract

The title of the graduation work: "Correlation and regression analysis of indicators of agricultural production"

In this final qualifying work, a multifactor model of an agricultural enterprise is built on the example of the Turkmenistan enterprise for the cultivation of cotton in the Lebap province of Turkmenistan. Estimates of the quality of the regression model are given, a quantitative analysis of the main indicators of the agricultural enterprise is carried out, the model is checked for adequacy, a program for the construction of a multivariate regression model is developed.

The following tasks were performed:

- a multifactor model for a productive attribute was built on the example of a specific agricultural enterprise, the model was checked for adequacy;
- developed software for building a multi-factor model.

For implementation, Python was used as a programming language using the libraries: numpy, statsmodels, math. A study was carried out using this program and the residues representing the forecast quality of the proposed model were obtained.

This undergraduate work consists of an explanatory note on 45 p., including 14 pictures.

Оглавление

Введение

От принятия оптимальных стратегических и тактических управленческих решений зависит состояние экономического процесса в дальнейшем. Использование основных методологических подходов и принципов применения аппарата эконометрического моделирования в процессе анализа позволяет оценить состояние, эффективность экономической системы а также спрогнозировать основные параметры и траектории ее развития под воздействием различных факторов и меняющихся условий функционирования.

Особенно сложно осуществить прогноз показателей в земледельческой отрасли, наиболее зависящей от климатических условий. Возможности современных математических методов и компьютерных технологий, конечно не могут решить все проблемы земледельческой отрасли, но позволяют создать картину функционирования отрасли на протяжении нескольких лет в виде чисел, функциональных зависимостей, схем, диаграмм…

Объектом работы является моделирование стохастических процессов на основе корреляционного и регрессионного анализа.

Предмет работы – методы регрессионного и корреляционного анализа.

Целью работы является построение многофакторной модели сельскохозяйственного предприятия для осуществления количественного анализа основных показателей сельскохозяйственного предприятия

Для достижения поставленной цели необходимо было выполнить следующие задачи:

- Построить многофакторную модель для результативного признака на примере конкретного сельскохозяйственного предприятия, проверить модель на адекватность;
- Разработать программное обеспечение для построения модели.

Глава 1. Основные понятия теории корреляционно-регрессионного анализа

1.1 Понятие о многомерном корреляционном анализе.

«Экономические явления часто адекватно описываются при помощи многофакторных моделей.

Пусть имеется набор случайных величин $(X_1, X_2, ..., X_i, ..., X_j, ..., X_p)$, которые имеют нормальное распределение. В этом случае матрицу

$$Q_p = \begin{pmatrix} 1 & p_{12} & \dots & p_{1p} \\ p_{21} & 1 & \dots & p_{2p} \\ \dots & \dots & \dots & \dots \\ p_{p1} & p_{p2} & \dots & 1 \end{pmatrix}, \quad (1.1)$$

составленную из парных коэффициентов корреляции $r_{ij} = (i, j = 1,2,..., p)$ определяемых по формуле

$$x = \frac{i = \sum_{i=1}^{m} x_i n_i}{n} \quad (1.2)$$

будем называть корреляционной. Основной задачей многомерного корреляционного анализа является оценка корреляционной матрицы Q_p по выборке. Эта задача решается определением матрицы выборочных коэффициентов корреляции:

$$q_p = \begin{pmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \dots & \dots & \dots & \dots \\ r_{p1} & r_{p2} & \dots & 1 \end{pmatrix}, \quad (1.3)$$

где $r_{ij} = (i, j = 1,2,..., p)$ определяется по формуле

$$r = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\sigma_x \sigma_y} \quad (1.4)$$

где $\sigma_x = \sqrt{\overline{x^2} \cdot \overline{x}^2}$, $\sigma_y = \sqrt{\overline{y^2} \cdot \overline{y}^2}$ – среднеквадратические отклонения.

В многомерном корреляционном анализе рассматривают две типовые задачи:

a) определение тесноты связи одной из переменных с совокупностью остальных $(p-1)$ переменных, включенных в анализ;
b) определение тесноты связи между переменными при фиксировании или исключение влияния остальных q переменных, где $q \le (p-2)$.

Эти задачи решаются с помощью множественных и частных коэффициентов корреляции» [12, с. 424-425].

1.2 Множественный коэффициент корреляции.

В отличие от парной корреляции и регрессии, где исследуется взаимосвязь двух переменных, в множественном корреляционном и регрессионном анализе рассматриваются взаимосвязи многих показателей, что более соответствует экономическим реалиям, системным взаимосвязям в экономических процессах. В эконометрических исследованиях множественный корреляционный анализ, оставаясь вспомогательным инструментом, играет важную аналитическую роль при исследовании отдельных специфических проблем эконометрики, в то же время исследователь, применяя множественный корреляционный анализ, может получить дополнительную полезную информацию о взаимосвязях экономических показателей. В экономике значение опрошенного показателя часто развивается под влиянием не одного, а множества факторов. Взаимосвязь между несколькими экономическими показателями изучается с использованием множественных корреляций. При этом могут исследоваться две проблемы: 1) влияние на один какой-либо показатель совокупности

факторов; 2) анализ взаимосвязи между любыми двумя факторами, исключая влияние обоих других факторов. Множественный корреляционный анализ основывается на парной корреляции.

Пусть имеется n экономических показателей. Чтобы проводить множественный корреляционный анализ, парные коэффициенты корреляции вычисляются между каждой парой экономических показателей.

Математической мерой корреляции двух случайных величин является корреляционная взаимосвязь или коэффициент корреляции. Если изменение одной случайной величины не приводит к регулярному изменению другой случайной величины, но приводит к изменению другого статистического признака этой случайной переменной, то это отношение не считается коррелированным, хотя оно является статистическим.

Тесноту связи изучаемых явлений оценивает коэффициент парной корреляции r_{xy} для регрессии. Коэффициент парной корреляции представляет собой безразмерную величину, равную (1.4).

Также можно использовать следующие формулы для расчета коэффициента корреляции:

$$r = \frac{COV(X,Y)}{\sigma_x \sigma_y} = \frac{M((X - M(X)(Y - M(Y))))}{\sqrt{D_x}\sqrt{D_y}} = \frac{M(XY) - M(X)M(Y)}{\sqrt{D_x}\sqrt{D_y}},$$

$$r = \frac{n\sum (x_i - \bar{x})(y_i - \bar{y})}{n\sqrt{\sum \mathrm{x}_\mathrm{i}^2 - (\sum x_i)^2 n\sqrt{\sum \mathrm{y}_\mathrm{i}^2 - (\sum y_i)^2}}} = \frac{n\sum x_i y_i - (\sum x_i)(y_i)}{n\sqrt{\sum \mathrm{x}_\mathrm{i}^2 - (\sum x_i)^2 n\sqrt{\sum \mathrm{y}_\mathrm{i}^2 - (\sum y_i)^2}}} \qquad (1.5)$$

Для практических расчетов наиболее удобна последняя формула.

Свойства коэффициента корреляции:

1) $|r| \le 1$;

2)чем ближе $|r|$ к 1, тем связь будет более тесной, а чем ближе $|r|$ к нулю, тем связь слабее;

3)если $r = +1$ или $r = -1$, то связь между значениями X и Y функциональная;

4) если $r < 0$, то связь между признаками обратная; если $r > 0$, то связь прямая;
5) если $r = 0$, то линейной корреляционной зависимости между признаками X и Y нет. Однако, между ними возможна нелинейная корреляционная зависимость.

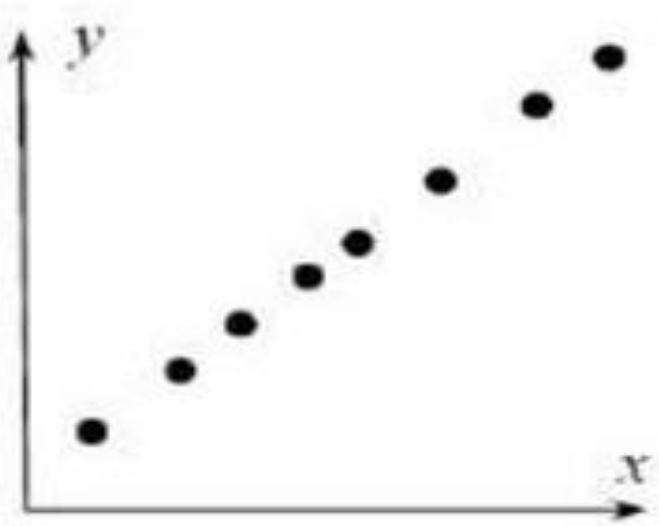


Рисунок 1.1 − График корреляционной зависимости.

На рисунке (1.1) представлено график корреляционной зависимости между признаками, когда их коэффициенты корреляции ближе к единицу.

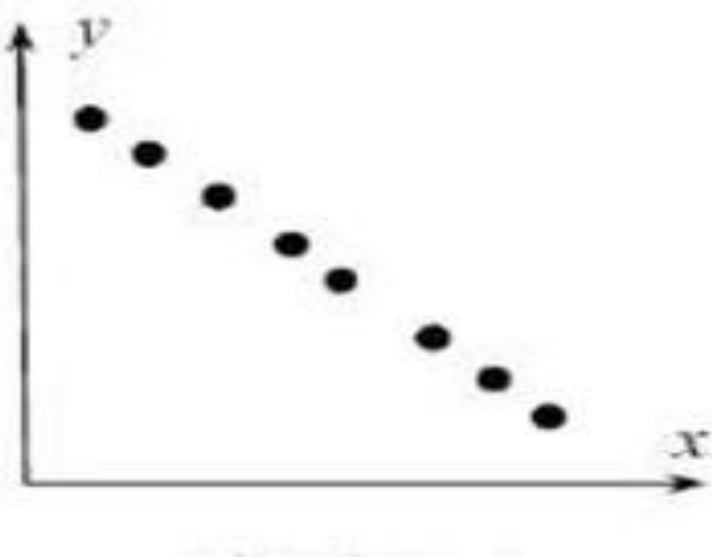


Рисунок 1.2 − График корреляционной зависимости.

На рисунке (1.2) представлен график корреляционной зависимости между признаками, когда их коэффициенты корреляции ближе к минус единицу.

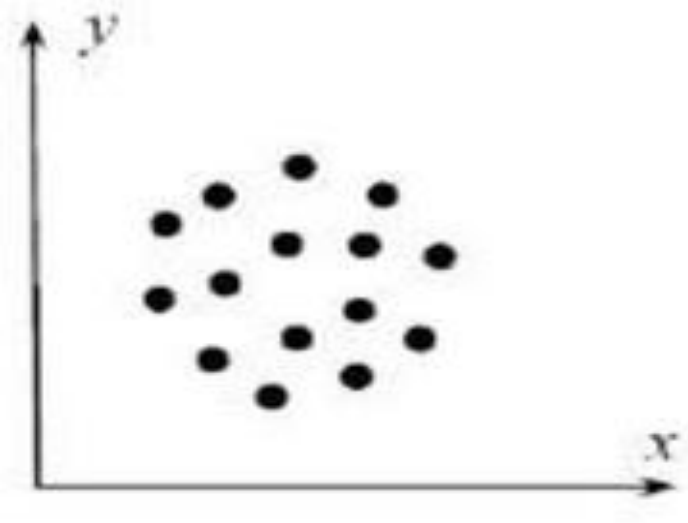


Рисунок 1.3 − График корреляционной зависимости

На рисунке (1.3) представлен график корреляционной зависимости между признаками, когда их коэффициенты корреляции ближе к нулю.

1.3 Частный коэффициент корреляции

Коэффициенты частной корреляции отличается от простого линейного парного коэффициента корреляции тем, что он измеряет парную корреляцию соответствующих переменных (y и X_i) при условии, что влияние остальных переменных (X_j) на них исключено.

«Выборочным частным коэффициентом корреляции (или просто частным коэффициентом корреляции) между переменными X_i или X_j при фиксированных значениях остальных $(p-2)$ переменных называется выражение

$$r_{ij.1,2,\ldots,p} = \frac{-q_{ij}}{\sqrt{q_{ii}q_{jj}}} \quad (1.6)$$

где q_{ii} и q_{jj} – алгебраические дополнение элементов r_{ij} и r_{jj} матрицы выборочных коэффициентов корреляции (1.3) а r_{ij} определяются по формулам (1.4).

В частности, в случае трёх переменных (n=3) из формулы (1.6),

следует, что

$$r_{ij.k} = \frac{r_{ij} - r_{ik} r_{jk}}{\sqrt{(1-r_{ik}^2)(1-r_{jk}^2)}} \qquad (1.7)$$

Поясним полученную формулу. Предположим, что имеется обычная регрессионная модель $x_i = \beta_0 + \beta_1 x_j + \beta_2 x_k + \varepsilon_i$ и необходимо оценить корреляцию между зависимой переменной X_i и объясняющей переменной X_j при исключении (элиминирование) влияние другой объясняющей переменной X_k . С этой целью найдем уравнение парной регрессии X_i по $X_k (\hat{x}_i = b_0 + b_1 x_k)$ и X_j по $X_k (\hat{x}_i = b'_0 + b'_1 x_k)$, а затем удалим влияние переменной X_k ,в остатки $e_{x_i} = x_i - \hat{x}_i$ и $e_{x_j} = x_j - \hat{x}_j$. Очевидно, что коэффициент корреляции между остатками e_{x_i} и e_{x_j} будет отражать тесноту частной корреляции между переменными X_i и X_j при исключении влияние переменной X_k . Можно показать, что найденный по формуле (1.4), обычный коэффициент корреляции между остатками e_{x_i} и e_{x_j} равен частному коэффициенту корреляции, $r_{ij.k}$ определенному по формуле (1.7)» [12, с. 426].

Коэффициент частный корреляции $r_{ij.1,2,\ldots,p}$, как и коэффициенты парной корреляции r_{ij} может принимать значение от минус единицы до единицы. Кроме того, $r_{ij.1,2,\ldots,p}$, вычисленный на основе выборки объёма n, имеет такое же распределение, как и r_{ij}, вычисленный по $n' = n - p + 2$ наблюдениям.

Поэтому значимость коэффициента частного корреляции $r_{ij.1,2,\ldots,p}$ оценивают так, же как и обычного коэффициента корреляции r, но при этом полагают $n' = n - p + 2.$

1.4 Основы регрессионного анализа

В регрессионном анализе рассматривается односторонняя зависимость случайной переменной Y от одной (или нескольких) неслучайной независимой переменной X.

Такая зависимость Y от X называют регрессионной. Она может быть представлена в виде модельного уравнения регрессии Y по X

X - факторный признак, объясняющая или экзогенная переменная;

Y - результативный признак, объясняемая или эндогенная переменная.

«В силу воздействия неучтенных случайных факторов и причин отдельные наблюдения переменной Y будут в большей или меньшей мере отклоняться от функции регрессии $\varphi(x)$. В этом случае уравнение взаимосвязи двух переменных (парная регрессионная модель) может быть представлено в виде:

$$Y = \varphi(X) + \varepsilon, \tag{1.8}$$

где ε - случайная переменная (случайный член), характеризующая отклонение от функции регрессии» [12, с. 439].

В случае линейного регрессионного анализа функция $\varphi(X)$ линейна относительно оцениваемых параметров:

$$M_x(Y) = \beta_0 + \beta_1 \cdot x. \tag{1.9}$$

Для оценки параметров β_0 и β_1 линейной функции регрессии (1.9) берется выборка, содержащая n пар значений переменных (x_i, y_i), где $i = 1,2,...,n$. В этом случае линейная парная регрессионная модель имеет вид:

$$y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i. \tag{1.10}$$

Уравнение

$$\widehat{y}_x = b_0 + b_1 \cdot x \tag{1.11}$$

является оценкой по выборке уравнения регрессии (1.10).

В парной регрессии выбор вида математической функции $\hat{y}_x = f(x)$ может быть осуществлен тремя методами:

- графическим;

- аналитическим, т.е. исходя из теории изучаемой взаимосвязи;

- экспериментальным.

При изучении зависимости между признаками графический метод подбора вида уравнения регрессии достаточно нагляден. Используя данные об индивидуальных значениях признака-фактора и соответствующих ему значениях результативного признака, можно построить в прямоугольных координатах точечный график, который называют «полем корреляции». По виду корреляционного поля можно определить имеется ли связь между признаками. Если связь между признаками есть и с ростом факторного признака результативный имеет тенденцию возрастать, то связь между признаками называется прямой, в обратном случае связь называется обратной.

Для того чтобы прогнозировать результативный признак, связь между признаками моделируется с помощью уравнения, которое называется уравнением регрессии.

Различают линейные и нелинейные регрессии.

1.5 Множественный регрессионный анализ

Парная регрессия на практике встречается довольно редко, более широкое применение получила множественная регрессия. При построении, например, модели зависимости роста товарооборота в какой-либо торговой организации от количества обслуживающего персонала исследователь предполагает, что в каждой организации одного типа (розничный магазин, оптовый склад) влияние на товарооборот таких факторов, как набор реализуемых товаров, их цена, торговая площадь, ее оформление, время

работы, одинаково. Однако данное предположение далеко не всегда справедливо. В связи с этим возникает задача исследования зависимости одной зависимой переменной У от нескольких объясняющих переменных $X_1, X_2,,,,,X_p$. Эта задача решается с помощью множественного регрессионного анализа.

При отборе факторов существуют определенные правила, выполнения которых необходимо, иначе оценки параметров уравнения и оно само будут недостоверными:

− факторы должны быть количественно измеримыми (допустим путем ранжирования), наличие или отсутствие свойства осуществляется приписыванием значений 0 или 1;

− каждый фактор должен быть тесно связан с результатом;

− факторы не должны быть тесно связаны между собой (не должны быть интеркоррелированы между собой), коэффициент корреляции между результативным и факторным признаками должен быть больше, чем коэффициент корреляции рассматриваемого фактора с остальными факторами.

«Основные предпосылки регрессионного анализа:

1. Зависимая переменная y_i (или возмущения ε_i) есть величина случайная, а объясняющая переменная x_i − величина неслучайная.

2. Математическое ожидание возмущение ε_i равно нулю:

$$M(\varepsilon_i) = 0. \qquad (1.12)$$

3. Дисперсия зависимой переменной y_i (или возмущения ε_i) постоянная для любого i:

$$D(\varepsilon_i) = \sigma^2. \qquad (1.13)$$

4. Переменные y_i и y_j (или возмущения ε_i и ε_j) не коррелированны:

$$M(\varepsilon_i, \varepsilon_j) = 0 \ (i \neq j). \qquad (1.14)$$

5. Зависимая переменная y_i (или возмущения ε_i) есть нормально распределенная случайная величина» [12, с. 440].

«Обозначим i-е наблюдение переменной y_i, а объясняющих переменных $-x_{i1}, x_{i2},..., x_{ip}$. Тогда модель множественной линейной регрессии можно представить в виде:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i, \quad (1.15)$$

где $i = 1,2,...,n$, а ε_i удовлетворяет приведенным выше предпосылкам 2–4.

Включение в регрессионную модель новых объясняющих переменных усложняет полученные формулы и вычисления. Это приводит к целесообразности использования матричных обозначений. Матричное описание регрессии облегчает как теоретические концепции анализа и так необходимые рачетные процедуры» [12, с. 454-455].

«Введем обозначения: $Y = (y_1, y_2,..., y_n)'$ –матрица-столбец, или вектор, значений зависимой переменной размера n;

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & ... & x_{1p} \\ 1 & x_{21} & x_{22} & ... & x_{2p} \\ ... & ... & ... & ... & ... \\ 1 & x_{n1} & x_{n2} & ... & x_{np} \end{pmatrix} \quad (1.16)$$

–матрица значений объясняющих переменных, или матрица плана размера $n \times (p+1)$ (обращаем внимание на то, что в матрицу X дополнительный введен столбец, все элементы которого равны 1, т.е. условно полагается, что в модели y_i свободный член β_0 умножается на фиктивную переменную x_{i0}, принимающую значение 1 для всех $i: x_{i0} \equiv 1 (i = 1,2,...,n)$);

$\beta = (\beta_0, \beta_1,..., \beta_p)'$ – матрица-столбец, или вектор, параметров размера $(p+1)$;

$\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)'$ −матрица-столбец, или вектор, возмущений (случайных ошибок остатков) размера n.

Тогда в матричной форме модель y_i примет вид:

$$Y = Xb + \varepsilon. \tag{1.17}$$

Оценкой этой модели по выборке является уравнение:

$$Y = Xb + e, \tag{1.18}$$

где $b = (b_0, b_1, ..., b_p)', e = (e_1, e_2, ..., e_n)$.

Для оценки вектора неизвестных параметров β применим метод наименьших квадратов. Так как произведение транспонированной матрицы e'на саму матрицу e:

$$e'e = (e_1, e_2, ..., e_n)\begin{pmatrix} e_1 \\ e_2 \\ ... \\ e_n \end{pmatrix} = \sum_i^n e_i^2 , \tag{1.19}$$

то условие минимизации остаточной суммы квадратов запишется в виде» [12, с. 455]:

$$S = \sum_{i=1}^{n} (y_{x_i} - y_i)^2 = \sum_{i=1}^{n} e_i^2 = e'e = (Y - Xb)' + (Y - Xb) \to \min \tag{1.20}$$

«Учитывая, что при транспонировании произведения матриц получается произведение транспонированных матриц, взятых в обратном порядке, т.е. $(Xb)' = b'X$, получим после раскрытия скобок:

$$S = Y'Y - b'X'Y - Y'Xb + b'X'Xb \tag{1.21}$$

Произведение $Y'Xb$ есть матрица размер $(1 \times n)[n \times (p+1)] \times [(p+1) \times 1] = (1 \times 1)$, т.е. величина скалярная, следовательно, оно не меняется при транспонировании:

$$Y'Xb = (Y'Xb)' = b'X'Xb \to \min. \tag{1.22}$$

На основании необходимого условия экстремума функции нескольких переменных $S(b_0, b_1, ..., b_p)$, представляющей S ,необходимо приравнять к нулю частные производные по этим переменным или в матричной форме – вектор частных производных

$$\frac{\partial}{\partial b} = \left(\frac{\partial S}{\partial b_0}, \frac{\partial S}{\partial b_1}, ..., \frac{\partial S}{\partial b_p} \right). \quad (1.23)$$

Для вектора частных производных доказаны следующие формулы:

$$\frac{\partial}{\partial b} = (b'c) = c, \frac{\partial}{\partial b} = (b'Ab) = 2Ab, \quad (1.24)$$

где b и c – вектор – столбец, а A –симметрическая матрица, в которой элементы, расположенные симметрично относительно главной диагонали, равны» [12, с. 456].

«Поэтому, полагая $c = X'Y$, а матрица $A = X'X$,найдем

$$\frac{\partial S}{\partial b} = -2X'Y' + 2X'Xb = 0 \quad (1.25)$$

оттуда получаем систему нормальных уравнений в матричной форме, для определения вектора b:

$$X'Xb = X'Y. \quad (1.26)$$

Найдем матрицы, входящие в это уравнение. Матрица $X'X$ представляет матрицу сумм первых степеней, квадратов и попарных произведений n наблюдений, объясняющих переменных:

$$X'X = \begin{pmatrix} 1 & 1 & ... & 1 \\ x_{11} & x_{12} & ... & x_{1p} \\ ... & ... & ... & ... \\ x_{1p} & x_{2p} & ... & x_{np} \end{pmatrix} \times \begin{pmatrix} 1 & x_{11} & ... & x_{1p} \\ 1 & x_{21} & ... & x_{2p} \\ ... & ... & ... & ... \\ 1 & x_{n1} & ... & x_{np} \end{pmatrix} = \begin{pmatrix} n & \sum x_{i1} & ... & \sum x_{ip} \\ \sum x_{i1} & \sum x_{i1}^2 & ... & \sum x_{i1}x_{ip} \\ ... & ... & ... & ... \\ \sum x_{ip} & \sum x_{i1}x_{ip} & ... & \sum x_{ip}^2 \end{pmatrix}. \quad (1.27)$$

Матрица $X'Y$ является вектор произведений n наблюдений объясняющих и зависимой переменных:

$$X'Y = \begin{pmatrix} 1 & 1 & ... & 1 \\ x_{11} & x_{12} & ... & x_{1p} \\ ... & ... & ... & ... \\ x_{1p} & x_{2p} & ... & x_{np} \end{pmatrix} \times \begin{pmatrix} y_1 \\ y_2 \\ ... \\ y_i \end{pmatrix} = \begin{pmatrix} \sum y_{ip} \\ \sum y_{i1}y_{ip} \\ ... \\ \sum y_{ip}^2 \end{pmatrix}. \quad (1.28)$$

В частном случае из рассматриваемого матричного уравнения $X'Xb = X'Y$ с учетом соотношений $X'X$ и $X'Y$ для одной объясняющей переменной (p = 1) нетрудно уже рассматриваемую систему нормальных уравнений, для не сгруппированных данных. Действительно, в этом случае матричное уравнение $X'Xb$ принимает вид:

$$\begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} \times \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum y_i x_i \end{pmatrix}, \quad (1.29)$$

оттуда непосредственно следует системе нормальных уравнений для не сгруппированных данных.

Для решения матричное уравнение $X'Xb$ относительно вектора оценок параметров b, необходимо ввести еще одну коэффициент регрессии для анализа множественной регрессии: матрица $X'X$ является неособенной, т.е. ее определитель не равен нулю.

Следовательно, ранг матрицы $X'X$ равен ее порядку, т.е. $r(X'X) = p+1$. Из матричной алгебры известно, что $r(X'X) = r(X)$, значит, $r(X) = p+1$, т.е. ранг матрицы плана равен числу ее столбцов.

Решением уравнения $X'Xb$ является вектор

$$b = (X'X)^{-1}X'Y \quad (1.30)$$

где $(X'X)^{-1}$— матрица, обратная матрице коэффициентов системы $X'Xb$, а $X'Y$ — матрица-столбец, или вектор, ее свободных членов.

Зная вектор b, выборочное уравнение множественной регрессии представим в виде:

$$y_{x0} = X'_0 b, \quad (1.31)$$

где y_{x0} — групповая (условная) средняя переменной Y при заданном векторе значений объясняющей переменной» [12, с. 457-458].

1.6 Определение доверительных интервалов для коэффициентов и функции регрессии.

Перейдем теперь к оценке значимости коэффициентов регрессии b_j и построению доверительно интервала для параметров регрессионной модели β_j $(j = 1,2,..., p)$.

«Оценка значимость коэффициента регрессии b_j можно проверить, если учесть, что статистика $(b_j - \beta_j)/S_{b_j}$ имеет t−распределение Стьюдента с $k = n - p - 1$ степенями свободы. Поэтому b_j значимо отличается от нуля на уровне значимости α, если $t = (b_j / S_{b_j}) \succ t_{1-\alpha;k}$, соответствующий $\gamma = (1-\alpha)\%$ −й доверительный интервал для параметра β_j есть

$$b_j - t_{1-\alpha;k} S_{b_j} \le \beta_j \le b_j + t_{1-\alpha;k} S_{b_j}. \qquad (1.32)$$

Наряду с интервальным оцениванием коэффициентов регрессии по условию весьма важным для оценки точности определения зависимой переменной (прогноза) является построение доверительного интервала для функции регрессии или для условного математического ожидания зависимой переменной $M_{x_0}(y)$, найденного в предположении, что объясняющие переменные $X_1, X_2,..., X_p$ приняли значение, задаваемые вектором $X'_0 = (1, x_{10}, x_{20},..., x_{p0})$. Обобщая соответствующие выражения на случай множественной регрессии, можно получить доверительный интервал для $M_{x_0}(y)$:

$$y_{x_0} - t_{1-\alpha;k} S_{y_{x_0}} \le M_{x_0}(Y) \le y_{x_0} + t_{1-\alpha;k} S_{y_{x_0}}, \qquad (1.33)$$

где y_{x_0} − групповая средняя, определяемая по уравнению регрессии,

$$S_{y_{x_0}} = s\sqrt{X'_0 (X'X)^{-1} X_0} \qquad (1.34)$$

−ее стандартная ошибка.

Доверительный интервал для индивидуальных значений зависимой переменной y_0^* примет вид» [12, с. 465]:

$$y_{x_0} - t_{1-\alpha;n-p-1} S_{y_0} \le y_0^* \le y_{x_0} + t_{1-\alpha;n-p-1} S_{y_0} \quad (1.35)$$

где

$$S_{y_0} = s\sqrt{1 + X'_0 (X'X)^{-1} X_0}\,. \quad (1.36)$$

1.7 Оценка взаимосвязи переменных. Проверка значимости уравнения множественной регрессии.

Одной из наиболее эффективных оценок адекватности регрессионной модели, мерой качества уравнения регрессии (мерой качества подгонки регрессионной модели к наблюденным значениям), является коэффициент детерминации:

$$R = \sqrt{\frac{Q_R}{Q}} = \sqrt{1 - \frac{Q_e}{Q}}\,, \quad (1.37)$$

где Q, Q_R и Q_e вычисляются по формулам:

$$Q = \sum_{i=1}^{n} (y_i - \bar{y})^2, \quad (1.38)$$

$$Q_R = \sum_{i=1}^{n} (y_{xi} - \bar{y})^2, \quad (1.39)$$

$$Q_e = \sum_{i=1}^{n} (y_i - y_{x_i})^2 \quad (1.40)$$

Получаем более подходящую формулу для R, которая не требует вычисления остатков e_i – и остаточной суммы квадратов

$$Q_e = \sum_{i=1}^{n} e_i^2\,. \quad (1.41)$$

В соответствии с Q, Q_R и Q_e

$$Q = \sum_{i=1}^{n} y_i^2 - 2y\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} \bar{y}^2 = \sum_{i=1}^{n} y_i^2 - 2\bar{y}(n\bar{y}) + n\bar{y}^2 = Y'Y - n\bar{y}^2 \quad (1.42)$$

$$(\text{ибо}\sum_{i=1}^{n} y_i^2 = (y_1, y_2, ..., y_n)\begin{pmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{pmatrix} = Y'Y). \quad (1.43)$$

С учетом S имеем

$$Q_e = Y'Y - bX'Y \quad (1.44)$$

(либо силу $X'Xb\ b'X'Xb = b'X'Y$).

Наконец,

$$Q_R = Q - Q_e = Y'Y - n\bar{y}^2 - (Y'Y - b'X'Y) = b'X'Y - n\bar{y}^2 \quad (1.45)$$

Таким образом,

$$R = \sqrt{\frac{Q_R}{Q}} = \sqrt{\frac{b'X'Y - n\bar{y}^2}{Y'Y - n\bar{y}^2}} \ . \quad (1.46)$$

«Коэффициент R является обобщением коэффициента корреляции в множественной модели. В зависимости от тесноты связи R может принимать значения от 0 до 1. Величина R^2 ,называемая множественным коэффициентом детерминации, показывает долю вариации зависимой переменной, обусловленную регрессией или изменчивостью объясняющей переменной.

Таким образом, множественный коэффициент детерминации R^2 можно рассматривать как меру качества уравнения регрессии, характеристику прогностической силы анализируемой регрессионной модели: чем ближе R^2 к единице, тем лучше регрессия описывают зависимость между объясняющими и зависимой переменными.

Недостатком коэффициента детерминации R^2 является то, что он вообще говоря, увеличивается при добавлении новых объясняющих переменных, хотя это не обязательно означает улучшение качества регрессионной модели. В этом смысле предпочтительнее использовать скорректированный коэффициент детерминации R^2 , определяемый по формуле

$$\hat{R}^2 = 1 - \frac{n-1}{n-p-1}(1-R^2), \qquad (1.47)$$

Из формулы следует, что чем больше число объясняющих переменных p, тем меньше $\hat{R}^2$ по сравнению с R^2. В отличие от R^2 скорректированный коэффициент $\hat{R}^2$ может быть уменьшаться, при введении в модель новых объясняющих переменных, не оказывающих существенного влияния на зависимую переменную. Однако даже увеличение скорректированного коэффициента детерминации $\hat{R}^2$ при введении в модель новой объясняющей переменной не всегда означает, что ее коэффициент регрессии значим (это происходит, как можно показать, только в случае, если соответствующее значение $t-$статистики больше единицы (по абсолютной величине), т.е. $|t|>1$. Другими словами, увеличение $\hat{R}^2$ еще не означает улучшение качества регрессионной модели. Если коэффициент детерминации R^2 известен, то F-критерий значимости уравнения регрессии можно записать в виде:

$$F = \frac{R^2(n-p-1)}{(1-R^2)p} > F_{\alpha;k_1;k_2}, \qquad (1.48)$$

где $k_1 = p$, $k_2 = n-p-1$, ибо в уравнении множественной регрессии, вместе со свободным членом оценивается $m = p+1$ параметров» [12, с. 469-470].

Табличные значение F-критерия, это максимальная величина отношения дисперсий, при заданном уровне значимости α. Вычисленное значение F-статистики признается достоверном (отличным от единицы), если оно больше табличного. В этом случае нулевая гипотеза H_0 об отсутствие связи признаков отклоняется и делается вывод о существенности этой связи.

Если же величина F-статистики окажется меньше табличной, то нулевая гипотеза не может быть отклонена без серьезного риска сделать неправильный вывод о наличии связи.

1.8 Мультиколлинеарность.

Мультиколлинеарность относится к высокой взаимной корреляции объясняющих переменных. Мультиколлинеарность может проявляться в функциональной (явной) и стохастической (скрытой) формах.

«При функциональной форме мультиколлинеарности. по крайней мере, одна из парных связей между объясняющими переменными является линейной функциональной зависимостью. В этом случае матрица $X'X$ особенная, так как содержит линейно зависимые векторы-столбцы и ее определитель равен нулю, т. е. нарушается предпосылка 6-го регрессионного анализа. Это приводит к невозможности решения соответствующей системы нормальных уравнений и получения оценок параметров регрессионной модели.

Однако в экономических исследованиях мультиколлинеарность чаще проявляется в стохастической форме, когда между, хотя бы двумя объясняющими переменными существует тесная корреляционная связь. Матрица $X'X$ в этом случае является неособенной, но ее определитель очень мал» [11, с. 108].

«В то же время вектор оценок b и его ковариационная матрица K в соответствии с формулами $b=(X'X)^{-1}X'Y$ и $K=\sigma^2(X'X)^{-1}$ пропорциональны обратной матрице $(X'X)^{-1}$, а значит, их элементы обратно пропорциональны величине определителя $|X'X|$. В результате получаются значительные средние квадратические отклонения (стандартные ошибки) коэффициентов регрессии $b_0, b_1, ..., b_p$ и оценка их значимости по t −критерию не имеет смысла, хотя в целом регрессионная модель может оказаться значимой по F −критерию.

Оценки становятся очень чувствительными к незначительному изменению результатов наблюдений и объема выборки. Уравнения регрессии в этом случае, как правило, не имеют реального смысла, так как некоторые из

его коэффициентов могут иметь неправильные с точки зрения экономической теории знаки и неоправданно большие значения.

Другой подход состоит в исследовании матрицы $X'X$. Если определитель матрицы $|X'X|$ либо ее минимальное собственное значение близки к нулю (например, одного порядка с накапливающимися ошибками вычислений), то это говорит о наличии мультиколлинеарности.

Для устранения или уменьшения мультиколлинеарности используется ряд методов. Самый простой из них (но далеко не всегда возможный) состоит в том, что из двух объясняющих переменных, имеющих высокий коэффициент корреляции (больше 0,8), одну переменную исключают из рассмотрения. При этом, какую переменную оставить, а какую удалить из анализа, решают в первую очередь на основании экономических соображений. Если с экономической точки зрения ни одной из переменных нельзя отдать предпочтение, то оставляют ту из двух переменных, которая имеет большой коэффициент корреляции с зависимой переменной» [11, с. 109-110].

В первой главе ВКР были рассмотрены и изучены темы: многомерный корреляционный анализ на основе парного, частного коэффициентов корреляции; многомерный регрессионный анализ; исследования на присутствие в модели мультиколлениарности.

Глава 2 Построение многофакторной модели сельскохозяйственного предприятия

2.1 Корреляционный анализ

«При построении регрессионной модели, в качестве результативного признака (объясняемой переменной Y) выбран показатель объемов производства в сопоставимых ценах. Выбор факторных признаков в работе (объясняющих переменных x_i) обусловлен теорией Жана Батиста Сея, согласно которой в общественном производстве взаимодействуют три главных фактора производства: труд, капитал и земля. При этом каждый из факторов несет в себе, как интенсивную, так и экстенсивную составляющие» [8].

Таблица 2.1 − Факторные признаки сельскохозяйственного предприятия

Факторные признаки	Составляющие факторных признаков	Показатели
Труд	Экстенсивные	Количество занятых человек
	Интенсивные	Производительность труда (выработка)
Земля	Экстенсивные	Посевная площадь
	Интенсивные	Урожайность
Капитал	Экстенсивные	Объем используемых основных фондов
	Интенсивные	Фондоотдача

Таблица 2.2−Значения результативного и факторных признаков

Отчетный период	Объем производства в сопоставимых ценах (манат)	Факторные признаки					
		Труд		Земля		Капитал	
		Экстенсивный	Интенсивный	Экстенсивный	Интенсивный	Экстенсивный	Интенсивный
		Количество занятых в отрасли (чел.)	Производительность труда (манат/чел.)	Посевная площадь (га)	Урожайность (ц/га)	Объем используемых основных фондов (манат)	Фондоотдача (процент %)

Продолжение таблицы 2.2

	y	x_1	x_2	x_3	x_4	x_5	x_6
2010	3142855	416	340	1180	1430	1403232	32
2011	2518666	442	326	1180	1200	1215863	28
2012	3140108	456	338	1200	1304	1435038	31
2013	3228636	430	403	1220	1329	1445628	34
2014	3114684	398	345	1220	1310	1331032	33
2015	3482757	422	429	1220	1356	1482419	35
2016	3487215	412	438	1200	1358	1456646	35
2017	2551891	368	348	1200	1270	937359	27
2018	2494900	340	340	1200	1290	622575	29
2019	2526515	344	352	1200	1210	1226931	34

Определим тесноту линейной взаимосвязи между объёмом производства (y) и факторными признаками:

- количеством занятых в отрасли человек (x_1);
- производительностью труда (x_2);
- посевной площадью (га) (x_3);
- урожайностью (x_4);
- объемом используемых основных фондов (x_5);
- фондоотдачей (x_6).

Чтобы определить тесноту связи между переменными вычислим множественные коэффициенты корреляции, которые являются обобщением парных коэффициентов корреляции. Предварительно вычислим по формуле (1.4) матрицу парной корреляции.

Таблица 2.3 −Матрица парной корреляции

Результативные и факторные признаки	Объём производства	Количество занятых в отрасли человек	Производительность труда	Посевной площадь	Урожайность	Объём используемых основных фондов	Фондоотдачи

Продолжение таблицы 2.3

		y	x_1	x_2	x_3	x_4	x_5	x_6
Объём производства	y	1	0,58	0,74	0,44	0,76	0,79	0,75
Количество занятых в отрасли человек	x_1	0,58	1	0,18	-0,04	0,29	0,74	0,16
Производительность труда	x_2	**0,74**	0,18	1	0,53	0,43	0,48	0,72
Посевной площадь	x_3	0,44	-0,04	0,53	1	0,13	0,20	0,52
Урожайность	x_4	**0,76**	0,29	0,43	0,13	1	0,42	0,45
Объём используемых основных фондов	x_5	**0,79**	**0,74**	0,48	0,20	0,42	1	0,70
Фондоотдачи	x_6	**0,75**	0,16	**0,72**	0,52	0,45	**0,70**	1

Если коэффициенты парной корреляции имеют большие значения ($r_{y/x_i} > 0,7$), то согласно шкале Чеддока, между ними есть высокая корреляционная зависимость.

Коэффициенты парной корреляции результативного признака – объемов производства (y) с четырьмя факторными признаками:

- производительностью труда (x_2);
- урожайностью (x_4);
- объемом используемых основных фондов (x_5);
- фондоотдачей (x_6),

имеют довольно большие значения (> 0,7), это говорит о достаточно высокой корреляционной зависимости результативного признака с перечисленными факторными признаками. Влияние количества занятых в отрасли человека (x_1) и размеров посевной площади (x_3) на объемы производства значительно ниже.

Коэффициент частной корреляции отличается от простого линейного парного коэффициента корреляции тем, что он измеряет парную корреляцию

соответствующих переменных (y и x_i) при условии, что влияние остальных переменных (x_j) на них исключено.

На основании коэффициентов частной корреляции можно сделать вывод об обоснованности включения переменных в регрессионную модель. Если значение коэффициента частной корреляции является низким или незначительным, то это означает, что связь между этим фактором и результативным признаков либо очень слабая, либо полностью отсутствует и поэтому может быть исключена из модели (при низких значениях не имеет смысла вводить фактор). Предварительно вычислим частные коэффициенты корреляции по формуле (1.7).

В таблице (2.4) представлены частные коэффициенты корреляции результативного признака объёмов производства и факторных признаков.

Таблица 2.4 − Частные коэффициенты корреляции

Результативные и факторные признаки		Частные коэффициенты корреляции при неизменных значениях факторов					
Объем производства		Количество занятых в отрасли человек	Производительность труда	Посевной площадь	Урожайность	Объём используемых основный фондов	Фондоотдача
	y	x_1	x_2	x_3	x_4	x_5	x_6
Количество занятых в отрасли человек	x_1	1,00	**0,79**	0,57	**0,76**	0,65	**0,81**
Производительность труда	x_2	0,68	1,00	0,09	**0,73**	**0,74**	0,45
Посевной площадь	x_3	0,67	0,66	1,00	**0,79**	**0,79**	0,66
Урожайность	x_4	0,58	**0,70**	0,54	1,00	**0,80**	**0,70**
Объём используемых основный фондов	x_5	0,0001	0,66	0,46	**0,78**	1,00	0,44
Фондоотдача	x_6	**0,70**	0,43	0,09	**0,72**	0,56	1,00

Коэффициент частной корреляции, результативного признака объёмов производства (y) с факторным признаком фондоотдачей (x_6) при неизменном значении количества занятых в отрасли человек (x_1), имеет значение больше

0,7. Следовательно, согласно шкале Чеддока, перечисленные признаки между собой имеют высокую корреляционную зависимость.

Обратим внимание на то, что при неизменном значении производительности труда (x_2) между объёмом производства (y) и количеством занятых в отрасли человек (x_1), а также между объёмом производства (y) и урожайностью (x_4) значения частных коэффициентов корреляции больше 0,7. Соответственно при неизменном значении урожайности (x_4) между объёмом производства (y) и количеством занятых в отрасли человек (x_1), между объёмом производства (y) и производительность труда (x_2), между объёмом производства (y) и посевной площадью (x_3), между объёмом производства (y) и объёмом используемых основной фондов (x_5), а также между объёмом производства (y) и фондоотдачей (x_6) значения частных коэффициентов корреляции больше 0,7. При неизменном значении объёма используемых основных фондов (x_5) между объёмом производства (y) и производительности труда (x_2), между объёмом производства (y) и посевной площади (x_3), а также между объёмом производства (y) и урожайностью (x_4) значения частных коэффициентов корреляции больше 0,7. При неизменном значении фондоотдачи (x_6) между объёмом производства (y) и количеством занятых в отрасли человека (x_1), а также между объёмом производства (y) урожайностью (x_4) значения частных коэффициентов корреляции больше 0,7.

Корреляционный анализ дает основание исключить из модели экстенсивные переменные - количество занятых в отрасли человека (x_1) и посевную площадь (x_3).

2.2 Построение регрессионной модели

На основании метода наименьших квадратов построено уравнение регрессии.

$$\hat{y} = -8800624{,}19 + 3175{,}64x_1 + 2248{,}58x_2 + 4512{,}77x_3 + 2493{,}70x_4 + 0{,}16x_5 + 2460827{,}19x_6 \quad (2.1)$$

В (2.1) результативный признак (объясняемая переменная) (y) − объем производства в сопоставимых ценах в усл.д.е.

Факторные признаки (объясняющие переменные) в (2.1):

x_1 - количество занятых в отрасли человек;

x_2 - производительность труда (усл.д.е./чел);

x_3 - посевная площадь (га);

x_4 - урожайность(усл.д.е./га);

x_5 - объем используемых основных фондов (усл.д.е.)

x_6 - фондоотдача (усл.д.е./ усл.д.е).

Известно, что одной из наиболее эффективных оценок адекватности регрессионной модели, является показатель качества является коэффициент детерминации R^2.

Для модели (2.1) получено значение коэффициента детерминации R^2 =0,972, что свидетельствует о том, что вариация исследуемой зависимой переменной Y (объемов производства в сопоставимых ценах) на 97,2% объясняется изменчивостью включенных в модель (2.1) объясняющих переменных.

Для оценки мультиколлениарности модели (2.1) рассмотрим матрицу парной корреляции (табл.2.3).

Обратим внимание на то, что между факторными признаками количеством занятых в отрасли человек (x_1) и объем используемых основных фондов (x_5), между производительностью труда (x_2) и фондоотдачей(x_6)

парные коэффициенты корреляции больше 0,7, а также между объемом используемых основных фондов (x_5) и фондоотдачей(x_6) парные коэффициенты корреляции также больше 0,7. Кроме того, значение определителя матрицы парных коэффициентов корреляции Δ довольно мало: $\Delta = 0{,}00054$. Все это говорит о присутствии модели в мультиколлениарных факторов, следовательно, оценки параметров модели ненадежны. Подтвердим это предположение оценкой статистики Фаррара-Глоубера. Вычислим наблюдаемое значение статистики Фаррара-Глоубера по формуле: $FG = -(n-1-\frac{1}{6}(2\cdot p+5))\ln|\Delta| = 46{,}39$, где $n = 10$ - количество наблюдений; $p = 6$ - количество факторных признаков. Фактическое значение критерия FG сравним с табличным значением критерия χ^2=25,0 с 15 степенями свободы и на уровне значимости $\alpha = 0{,}05$. Так как $|FG| > \chi^2$ (46,4>25,0), то это свидетельствует о том, что в массиве объясняющих переменных существует мультиколлинеарность.

Из приведенных выше формальных исследований, а также на основании корреляционного анализа, исключим из модели экстенсивные переменные - количество занятых в отрасли человека(x_1) и посевную площадь (x_3). Новая модель примет вид:

$$\hat{y} = -2341886{,}15 + 3351{,}95x_2 + 2516{,}95x_4 + 0{,}66x_5 + 92608{,}27x_6 \qquad (2.2)$$

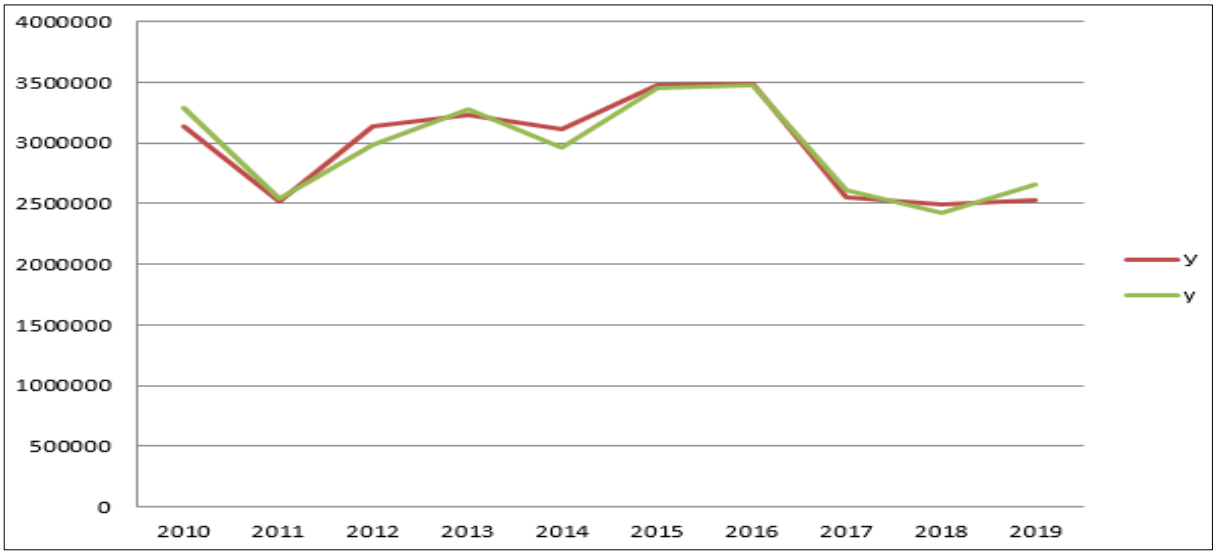


Рисунок 2.5 − Графики объёмов производства

На рисунке (2.5) представлены графики объёмов производства (Y) и (y), построенные на основе эмпирических данных и теоретических на основании уравнения (2.2)соответственно.

В таблице (2.5) указана матрица парной корреляции при исключении количества занятых в отрасли человек (x_1) и размеров посевной площади(x_3).

Таблица 2.5 –Скорректированная матрица парной корреляции

Результативные и факторные признаки		Объём производства	Производительность труда	Урожайность	Объём используемых основных фондов	Фондоотдачи
		y	x_2	x_4	x_5	x_6
Объём производства	y	1	0,74	0,76	0,79	0,75
Производительность труда	x_2	**0,74**	1	0,43	0,48	0,72
Урожайность	x_4	**0,76**	0,43	1	0,42	0,45
Объём используемых основных фондов	x_5	**0,79**	0,48	0,42	1	0,70
Фондоотдачи	x_6	**0,75**	**0,72**	0,45	**0,70**	1

В результате значение определителя скорректированной матрицы парных коэффициентов корреляции Δ для модели (2.2) увеличится до значения $\Delta = 0{,}012$. Наблюдаемое значение статистики Фаррара−Глоубера $FG = 2{,}49$ меньше табличного значения критерия $\chi^2 = 18{,}3$, т.е. статистика Фаррара−Глоубера не подтверждает мультиколлениарность модели (2.2) на 5% уровне значимости. Обратим внимание на значение коэффициента детерминации: $R^2 = 0{,}933$. Таким образом, расчетные параметры модели характеризует на 93,3% долю вариации зависимой переменной, обусловленной регрессией.

В таблице (2.6) проведена проверка модели (2.2) на адекватность значений коэффициентов регрессии с помощью критерия Стьюдента на уровне значимости α=0,2 с числом степеней свободы k=5.

Таблица 2.6. −Значение t−наблюдения и значение t−критерия Стьюдента

	Производительность труда	Урожайность	Объём используемых основных фондов	Фондоотдачи
	x_2	x_4	x_5	x_6
t-наблюдения	1,98	3,20	2,71	10,27
t-критерия	1,47	1,47	1,47	1,47

Сравнивая значение t-наблюдения с значениями t-критерия по таблице (2.6) можно сказать, что все коэффициенты регрессии (2.2) значимы (гипотеза о равенстве нулю соответствующего коэффициента регрессии). Это говорит о том, что все рассматриваемые факторы имеют влияние на объем производства (y).

Из приведенных выше формальных исследований и теоретических оснований найдём доверительные интервалы для значимых коэффициентов регрессии.

Заметим что, если в границы доверительного интервала попадает нуль, т.е. нижняя граница отрицательна, а верхняя положительна, то оцениваемый параметр принимается равным нулю, так как он не может одновременно принимать и положительное и отрицательное значения.

Таблица 2.7 − Значение для доверительного интервала.

Факторные признаки модели		β_{min}	B_{max}	Значения коэффициентов регрессии
Производительность труда	x_2	858,4378986	5845,460698	3351,95
Урожайность	x_4	1358,012218	3675,892683	2516,95
Объём	x_5	0,300404841	1,017288023	0,66
используемых основных фондов				
Фондоотдачи	x_6	30910,11	154306,93	92608,30

По данным таблицы (2.7) можно сказать, что доверительный интервал (858,44; 5845,46) покрывает неизвестный параметр коэффициент k_2, с вероятностью p=0,8. Соответственно доверительной интервал (1358,01; 3675,89) покрывает неизвестный параметр коэффициент k_4, доверительной интервал (0,30; 1,01) покрывает неизвестный параметр коэффициент k_5 и доверительной интервал (30910,11; 154306,93) покрывает неизвестный параметр коэффициент k_6 также с вероятностями p=0,8.

Проведем оценку значимости уравнения регрессии (2.2) по критерию Фишера. На уровне значимости $\alpha = 0{,}05$, проверим нулевую гипотезу о равенстве нулю параметров регрессии (2.2). Значение F − статистики, вычислим по формуле

$$F = \frac{R^2(n-p-1)}{(1-R^2)\cdot p}, \qquad (2.3)$$

где $k_1 = p,\ k_2 = n - p - 1$.

В результате получим фактическое значение критерия $F = 277{,}05$, больше табличного $F_{\alpha,k_1,k_2} = 5{,}1$ ($F > F_{\alpha,k_1,k_2}$), следовательно, уравнение (2.2) значимо на 5% уровне, исследуемая зависимая переменная $У$ достаточно хорошо описывается включенными в регрессионную модель переменными x_2, x_4, x_5, x_6.

Регрессионный анализ подтверждает, что при увеличении производительности труда (x_2) (при неизменном x_4, x_5, x_6) на одного рабочего объём производства (y) увеличивается в средним на 3351,95 манат. При увеличении урожайности (x_4) (при неизменном x_2, x_5, x_6) на одного центнера с гектара объём производства (y) увеличивается в средним на 2516,95 манат. При увеличении объема используемых основных фондов (x_5) (при неизменном x_2, x_4, x_6) на одного маната объём производства (y) увеличивается в средним на 0,66 манат. При увеличении фондоотдачи (x_6) (при неизменных значениях x_2, x_4, x_5) на одного процента (%) объём производства (y) увеличивается в средним на 92608,27 манат.

Во второй главе на основе данных сельскохозяйственного предприятия по выращиванию хлопка в Лебапском велаяте Туркменистана строится многофакторная модель, приводятся оценки качества регрессионной модели, осуществляется количественный анализ основных показателей сельскохозяйственного предприятия.

Глава 3 Программное обеспечение для многофакторной модели

3.1 Описание алгоритма многофакторного модели

В ходе выполнения бакалаврской работы для многофакторной модели на языке программирования Python было разработано программное обеспечение.

Для разработки программного обеспечения многофакторной модели использовались следующие пакеты.

«Numpy−библиотека с открытым исходным кодом для языка программирования Python с возможностью поддержки многомерных массивов (включая матрицы), высокоуровневых математических функций, предназначенных для работы с многомерными массивами.

Statsmodels − является частью научного файла Python, посвященного анализу данных и науку о данных и статистике .

Matplotlib − библиотека на языке программирования Python для визуализации данных двумерной (2D) графикой (также поддерживается 3D графика).

Math модуль − набор функций для выполнения математических, тригонометрических и логарифмических операций» [13].

На выходе программы – наборы многофакторных моделей, их характеристики и графическое представление результатов.

На основе анализа набора задач и материалов, полученных из различных источников информации, был получен следующий формальный алгоритм для автоматического составления оптимальной многофакторной модели.

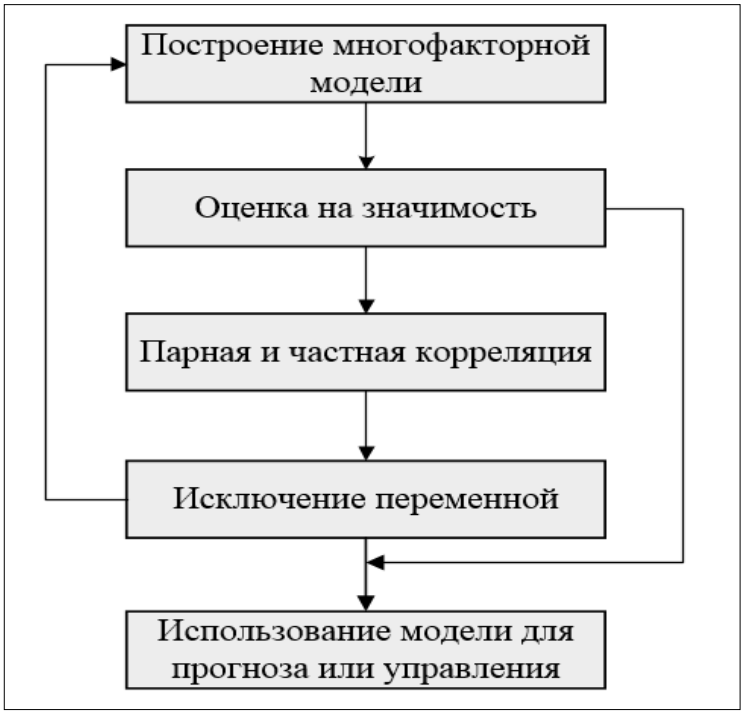


Рисунок 3.1 − Схема алгоритма многофакторного модели.

Опишем алгоритм подробно.

Шаг 1. На этом шаге вычисляем коэффициенты и строим уравнение регрессии, находим коэффициент детерминации, вычисляем значение F−статистики для уравнения регрессии. Далее вычисляем значения t−статистики, стандартные ошибки, p-вероятность и доверительные интервалы для коэффициентов регрессии.

Шаг 2. В результате второго шага проведем оценку значимость модели регрессии с помощью критерия Фишера.

Шаг 3. Вычисляем матрицу парных и частных коэффициентов корреляции. Эти матрицы показывают связь между факторами.

Шаг 4. Используя полученные результаты на предыдущем шаге матриц парной и частной корреляции, построим новую модель, исключив одну переменную из коллинеарных пар.

Шаг 5. Полученную многофакторную модель, в результате выполнения предыдущих шагов, используем для прогноза или управления.

3.2 Описание программного обеспечения

В первую очередь опишем программный код построения многофакторной модели.

```
def reg_m(y, x):
    ones = np.ones(len(x[0]))
    X = sm.add_constant(np.column_stack((x[0], ones)))
    for ele in x[1:]:
        X = sm.add_constant(np.column_stack((ele, X)))
    results = sm.OLS(y, X).fit()
    return results
print (reg_m(y, x).summary())
```

Рисунок 3.2 − Фрагмент кода для регрессионной модели

С помощью функция corrcoef опишем программный код для парных коэффициентов корреляции.

```
## Парная корреляция
corr=np.corrcoef(y, x)
print ("\nПарные коэффициенты корреляции \n", corr)
```

Рисунок 3.3 − Фрагмент кода парной корреляции

Опишем программный код для частного коэффициента корреляции.

```
def ch_k_k(x):
      r=x
      private_corr_coef=[[0]*7]*7
      for i in range(1,7):
            for j in range(1, 7):
                  if (i != j):
                        r[i][j] = round((x[i][j] - (x[i][0] * x[j][0])) / math.sqrt((1 - math.pow(x[i][0], 2)) * (1 - math.pow(x[j][0], 2))),3)
                        private_corr_coef[i][j]=r[i][j]
      print('\n')

      r[1:].clear()
      private_corr_coef=np.array(r)
      private_corr_coef.transpose()
      for i in range(7):
            private_corr_coef[0][i]=0
            private_corr_coef[i][0]=0
      print ("Частной коэффициент корреляции \n", private_corr_coef)
```

Рисунок 3.4 − Фрагмент кода частного коэффициента корреляции

Далее представим программный код для построения графиков.

```
fig, ax = plt.subplots()
ax.plot(x0, y, label='Объём производства (y) построеннных на основе эмпирических данных')
ax.plot(x0, y0, label='Объём производства (y) постраенных на основании уравнение регрессии')
ax.legend(loc='upper left')
ax.set_xlabel(r'Годы')
ax.set_ylabel(r'Объём производства y')
plt.show()
```

Рисунок 3.5 − Фрагмент кода построения графика

1.3 Проверка модели на адекватность

В качестве примера загрузим данные из таблицы (2.2). После того как загрузили, данные автоматически обрабатывается и строится многофакторная модель.

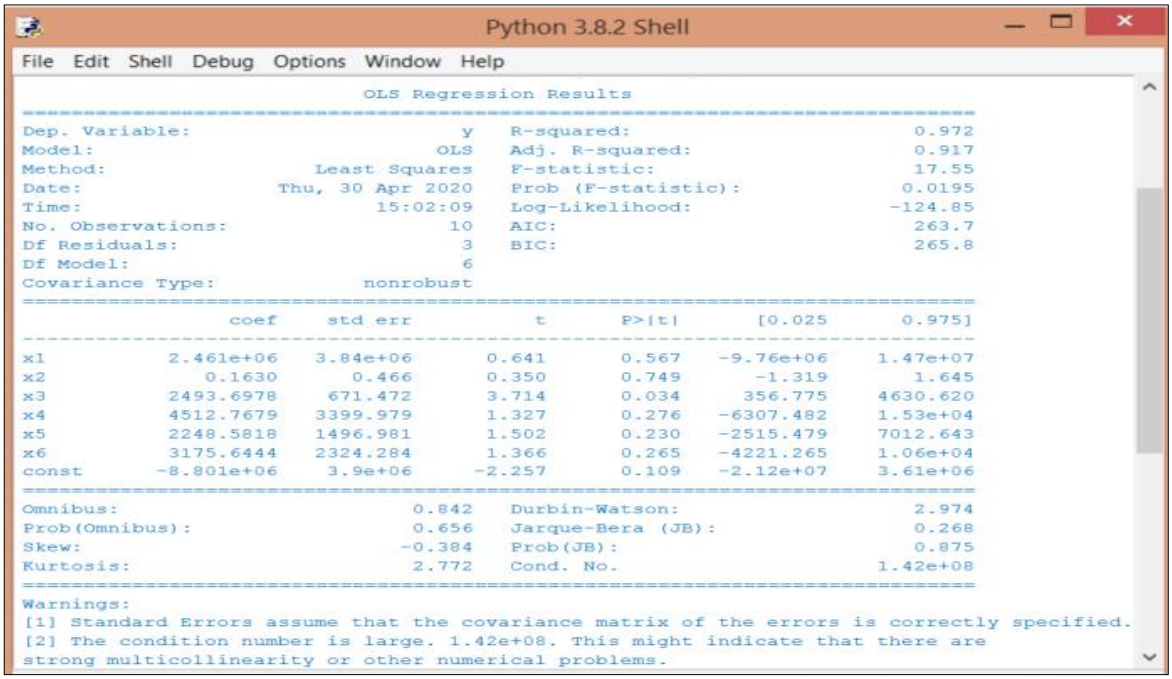


Рисунок 3.6 − Результат алгоритма построение многофакторной модели.

Как видно, что в этой модели существует мултикалениарность. Для того чтобы избавиться от мултикалениарности необходимо исключить одну переменную из коллинеарных пар. Для определения какую переменную исключить из коллинеарных пар, вычислим парный и частный коэффициенты корреляции.

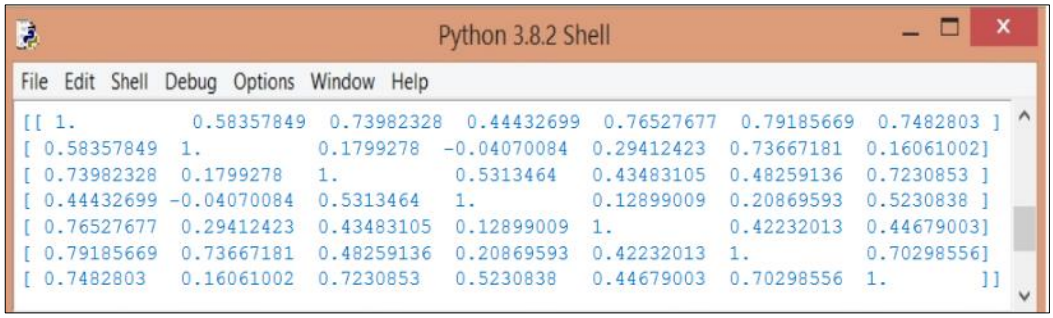


Рисунок 3.7 –Результат парной корреляции

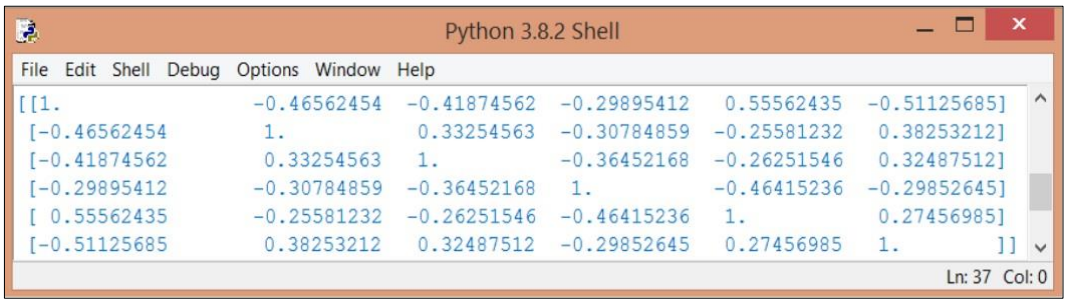
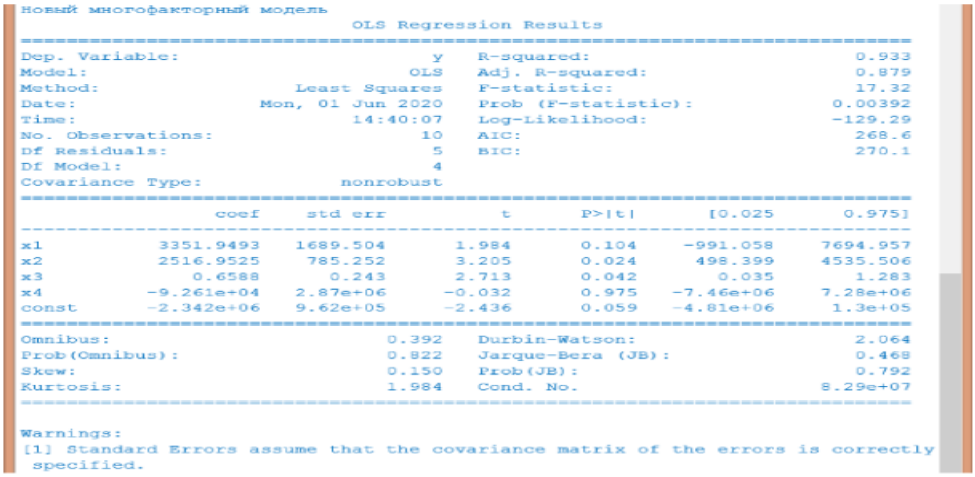


Рисунок 3.8 − Результат - фрагмент кода частного коэффициента корреляции

После того как мы получили результаты парных и частных коэффициентов корреляции, сравнивая их значения между собой, определяем какие переменные исключить из модели. В нашем случае исключим количество занятых в отрасли человек и посевную площадь. Построим новую многофакторную модель.



Новый многофакторный модель

OLS Regression Results

Dep. Variable:	y	R-squared:	0.933
Model:	OLS	Adj. R-squared:	0.879
Method:	Least Squares	F-statistic:	17.32
Date:	Mon, 01 Jun 2020	Prob (F-statistic):	0.00392
Time:	14:40:07	Log-Likelihood:	-129.29
No. Observations:	10	AIC:	268.6
Df Residuals:	5	BIC:	270.1
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	3351.9493	1689.504	1.984	0.104	-991.058	7694.957
x2	2516.9525	785.252	3.205	0.024	498.399	4535.506
x3	0.6588	0.243	2.713	0.042	0.035	1.283
x4	-9.261e+04	2.87e+06	-0.032	0.975	-7.46e+06	7.28e+06
const	-2.342e+06	9.62e+05	-2.436	0.059	-4.81e+06	1.3e+05

Omnibus:	0.392	Durbin-Watson:	2.064
Prob(Omnibus):	0.822	Jarque-Bera (JB):	0.468
Skew:	0.150	Prob(JB):	0.792
Kurtosis:	1.984	Cond. No.	8.29e+07

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Рисунок 3.9 Результат новой многофакторной модели

Кроме пошаговых результатов, также выводятся графики полученные на основе эмпирических данных и теоретических на основании уравнения регрессии.

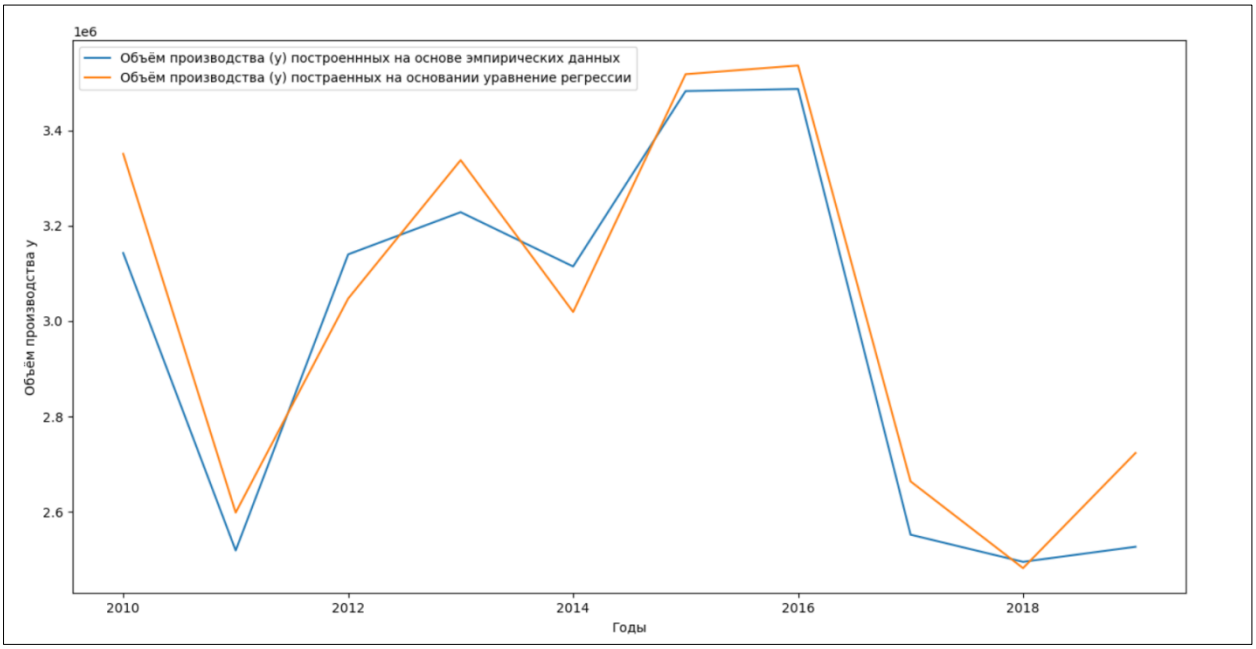


Рисунок 3.10 Графики, построенные на основе эмпирических данных и теоретических

В третьей главе был рассмотрен алгоритм построения многофакторной модели. Алгоритм выполняет следующие задачи: строит уравнение регрессии, проверяет на адекватность уравнение регрессии, вычисляет парные и частные коэффициенты корреляции, исключает одну переменную из коллинеарных пар и строит новую многофакторную модель.

Алгоритм многофакторного модели был реализован в среде разработки IDLE, с использованием языка программирование Python 3.8.

Заключения

Цель данной выпускной квалификационной работы состояла в построение многофакторной модели сельскохозяйственного предприятия для осуществления количественного анализа основных показателей. В результате были достигнуты все поставленные задачи:

- Построена многофакторная модель для результативного признака на примере конкретного сельскохозяйственного предприятия, проверена модель на адекватность.
- Разработано программное обеспечение для построения модели.

В первой главе рассмотрена основные понятия теории корреляционно – регрессионного анализа.

Во второй главе на основе данных сельскохозяйственного предприятия по выращиванию хлопка в Лебапском велаяте Туркменистана построена многофакторная модель, приводятся оценки качества регрессионной модели, осуществляется количественный анализ основных показателей сельскохозяйственного предприятия.

В третьей главе реализовано программное обеспечение прогнозирования на основе модели множественной регрессии с использованием языка программирования Python.

Представленный в работе пример построения и оценки качества регрессионной модели позволяет провести количественный анализ основных показателей сельскохозяйственного предприятия. В результате применения основных методологических подходов и принципов аппарата эконометрического моделирования появляется возможность получения множества разнообразных научно обоснованных вариантов перспективного развития предприятия, что позволяет руководителям составить прогноз экономического развития предприятия и принять наиболее оптимальные стратегические и тактические решения в процессе управления.

Список используемой литературы

Научная и методическая литература

1. Айвазян, С. А. Эконометрика. — М.:; ИНФРА-М, 2014.

2. Айвазян, С. А. Методы эконометрики : учебник / С. А. Айвазян, В. С. Мхитарян. — М.: Магистр; ИПФРА-М, 2010.

3. Гуриков С. Р. Основы алгоритмизации и программирования на Python : учеб. пособие — М. : ФОРУМ : ИНФРА-М, 2018. — 343 с.

4. Дайитбегов Д.М. Компьютерные технологии анализа данных в эконометрике: Монография / 3-е изд., испр. и доп. - М.: Вузовский учебник: НИЦ Инфра-М, 2013.

5. Доусон М. Программируем на Python. – СПб.: Питер, 2014. – 416 с.

6. Дроздова Е.М. Корреляционно-регрессионный анализ показателей сельскохозяйственного производства приморского края/ Дискуссия. -2013.- №11(41).- С.55-59;

7. Дайитбегов, Д.М. Компьютерные технологии анализа данных в эконометрике: Монография / Д.М. Дайитбегов. — М.: Вузовский учебник, НИЦ ИНФРА-М, 2013. — 587 с.

8. Дроздова Е.М. Корреляционно-регрессионный анализ показателей сельскохозяйственного производства приморского края/ Дискуссия. - 2013.- №11(41).- С.55-59;

9. Ковалев, Е. А. Теория вероятностей и математическая статистика для экономистов : учебник и практикум для бакалавриата, специалитета и магистратуры / Е. А. Ковалев, Г. А. Медведев ; под общ. ред. Г. А. Медведева. — 2-е изд., испр. и доп. — М. : Издательство Юрайт, 2019. — 284 с.

10. Касимова Т.М., Алиева Р.М. Оценка ресурсного потенциала сельскохозяйственных предприятий с помощью корреляционно-регрессионного анализа (на примере хозяйств Хунзахского района

республики Дагестан)/Фундаментальные исследования. - 2016. - № 4 (часть 1) - С. 166-169.

11. Кремер Н.Ш. Эконометрика/ Н.Ш. Кремер, Б.А. Путко. - М.: 2010.- 328с.

12. Кремер Н.Ш. Теория вероятностей и математическая статистика: Учебник для вузов/ 3-е изд., перераб. и доп. - М.: 2010 – 551 с.

13. Маккинли, У. Python и анализ данных / УэсМаккинли ; пер. с англ. А.А. Слинкина. - Москва : ДМК Пресс, 2015. - 482 с.

14. Новиков А. И. Эконометрика: Учебное пособие / 3-е изд., перераб. и доп. - М.: НИЦ ИНФРА-М, 2014. - 272 с.: 60x88 1/16.

15. Саммерфилд, М. Программирование на Python 3. Подробное руководство - М.: Символ-Плюс, 2011. - 608 с.

16. Симчера, В.М. Методы многомерного анализа статистических данных / В.М. Симчера. — М.: Финансы и статистика, 2018. — 400 с.

17. Тюрин, Ю.Н. Анализ данных на компьютере: Учебное пособие / Ю.Н. Тюрин, А.А. Макаров; Науч. ред. В.Э. Фигурнов. — М.: ИД ФОРУМ, 2017 — 368 с.

18. Уткин В.Б./Эконометрика, - 2-е изд. - М.:Дашков и К, 2017. - 564 с.

19. Чашкин, Ю.Р. Математическая статистика. Анализ и обработка данных: Учебное пособие / Ю.Р. Чашкин; Под ред. С.Н. Смоленский. — Рн/Д: Феникс, 2017. — 236 с.

Литература на иностранном языке

20. Box, George; Jenkins, Gwilym M.; Reinsel, Gregory C. (2016). Time Series Analysis: Forecasting and Control (Fifth ed.). ISBN 978-1-118-67502-1.

21. Jagannathan, R., Skoulakis, G. & Wang, Z. (2010), The analysis of the cross section of security returns, in Y. Aït-Sahalia& L. P. Hansen, eds, ‘Handbook of financial econometrics’, Vol. 2, Elsevier B.V., pp. 73–134.

22. Downey A., Elkner J., Meyers Ch. How to Think Like a Computer Scientist: Learning with Python. - Wellesley, Massachusetts: Green Tea Press, 2002.

23. Judge, G.G., R.C. Hill, W.E. Griffiths, H. L¨utkepohl and T.C. Lee (1985), The Theory and Practice of Econometrics (John Wiley and Sons: New York).

24. Bollerslev, T. & Wooldridge, J. M. (1992), 'Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances', Econometric Reviews 11(2), 143–172.