**Министерство науки и высшего образования Российской Федерации**

федеральное государственное бюджетное образовательное учреждение

высшего образования

Российский экономический университет имени Г.В. Плеханова

**Высшая школа кибертехнологий, математики и статистики**

# **Направление** Прикладная математика и информатика

**Профиль** Прикладная математика и информатика

**Кафедра** Математических методов в экономике

**Курсовая работа**

**по дисциплине**

**«Эконометрика и эконометрическое моделирование»**

**на тему**

Тема: «Эконометрическое моделирование инвестиций в основной фонд субъектов Российской Федерации»

|  |  |
| --- | --- |
|  | Работу выполнил:  студент 15.11Д-ПМПМИ03/21б группы  очной формы обучения  Тен Дмитрий Игорьевич  Научный руководитель:  д.э.н., профессор Тихомирова Т. М. |

Москва, 2024

**Аннотация**

Данная курсовая работа нацелена на построение эконометрической модели, удовлетворяющей всем требованиям. В данной работе будут разобраны такие темы, как: отбор признаков, свертывание данных, корреляционный анализ, тестирование и нивелирование мультиколлинеарности, робастно-статистическое оценивание, метод главных компонент, гетероскедастичность и прогнозирование.

При построении модели будут использованы реальные данные, полученные из открытых источников. В качестве объясняемого показателя были взяты инвестиции в основной капитал, в качестве признаковых переменных – социально-экономические показатели, описанные в первой главе. Показатели были отобраны по регионам Российской Федерации.

Планируется применить теоретические знания в сфере эконометрического моделирования на практических данных. В результате будет получена рабочая модель, с помощью которой можно будет провести анализ социально-экономические явления для регионов Российской Федерации.

***Ключевые слова*:** Эконометрическое моделирование, мультиколлинеарность, регионы России, инвестиции в основной капитал, робастное статистическое исследование, робастная регрессия, корреляционный анализ, многомерные данные.

**Abstract**

This coursework aims at building an econometric model that satisfies all the requirements. This paper will cover topics such as: feature selection, data convolution, correlation analysis, multicollinearity testing and leveling, robust statistical estimation, principal component method, heteroscedasticity and forecasting.

Real data obtained from public sources will be used in model building. Investments in fixed assets were taken as an explanatory indicator, and socio-economic indicators, described in the first chapter, were taken as sign variables. The indicators were selected by regions of the Russian Federation.

It is planned to apply theoretical knowledge in the field of econometric modeling on practical data. As a result, a working model will be obtained, with the help of which it will be possible to analyze socio-economic phenomena for the regions of the Russian Federation.

***Keywords***: Econometric modeling, multicollinearity, Russian regions, investment in fixed assets, robust statistical study, robust regression, correlation analysis, multivariate data.

**Содержание**

* **Глава 1. Многомерный корреляционный анализ. 2**

1. **Первичный анализ и обработка многомерных данных 2**
2. **Корреляционный анализ 9**
3. **Построение регрессионной модели. 10**

* **Глава 2. Эффект мультиколлинеарности 14**
* **Глава 3. Робастное статистическое оценивание 16**
* **Глава 4. Метод главных компонент 23**
  1. **Робастное оценивание с помощью метода главных компонент 23**
  2. **Применение метода главных компонент на факторных признаках 27**
  3. **Применение метода главных компонент на типичных данных 30**
* **Глава 5. Эффект гетроскедастичности остатков регрессионной модели 31**
  1. **Проверка модели на гетероскедастичность. 31**
  2. **Устранение гетероскедастичности 34**
* **Глава 6. Расчет прогнозных значений 37**
* **Заключение 38**
* **Источники 40**

**Введение**

Инвестиции в основной фонд представляют собой суммарные выделенные денежные средства, в данном случае определенному субъекту России, направленные на создание основных средств или реконструкцию существующих. Основными средствами считаются здания, транспортные средства и другие важные объекты. Данный показатель напрямую влияет на благополучие граждан определенного региона, в особенности на инфраструктуру населенных пунктов.

Анализ данного показателя важен для дальнейшего принятия решений в различных сферах деятельности. Прогноз инвестиций в основной капитал позволит строить планы развития регионов и распределения бюджета.

В основе предстоящего анализа лежит эконометрическое моделирование. Далее будут отобраны признаки для данного показателя, после чего будет произведен всесторонний анализ линейной регрессионной модели в рамках эконометрического моделирования.

**Глава 1. Многомерный корреляционный анализ**

**1.1. Первичный анализ и обработка многомерных данных.**

**Объект исследования:** Инвестиции в основной капитал по субъектам Российской Федерации.

**Перечень характеристик:**

1. **Удельный вес занятого населения (15-72 лет)** = численность занятых / численность экономически активного населения (%)
2. **Валовый региональный продукт, сведенный к ценам 2010 года** (рублей на душу населения), представляет собой валовую добавленную стоимость, созданную резидентами региона, и определяется как разница между выпуском и промежуточным потреблением
3. **Среднедушевые денежные доходы населения, сведенные к ценам 2010 года** (руб. / месяц) = годовой объем денежных доходов / (12 \* среднегодовая численность населения)
4. **Оборот средних организаций, сведенный к ценам 2010 года** (рублей на душу населения). Среднее предприятие — выручка от 800 млн. до 2 млрд. рублей в год, от 100 до 250 сотрудников.
5. **Численность фактически действующих индивидуальных предпринимателей** (на 100 тыс. человек)
6. **Общее количество преступлений** (на 100 тыс. человек)
7. **Индекс производительности труда** (%) = индекс физического объема ВВП / индекс совокупных затрат труда
8. **Число реализованных турпакетов населению** (на 100 тыс. человек)
9. **Оборот розничной торговли, сведенный к ценам 2010 года** (рублей на душу населения).
10. **Объем ликвидированных основных фондов за год по полному кругу организаций, сведенный к ценам 2010 года** (рублей на душу населения). Основные фонды - произведенные активы, подлежащие использованию неоднократно или постоянно в течение длительного периода, но не менее одного года, для производства товаров, оказания рыночных и нерыночных услуг, для управленческих нужд либо для предоставления другим организациям за плату во временное владение и пользование или во временное пользование.

**Инвестиции в основной капитал, сведенные к ценам 2010 года** – это совокупность затрат, направленных на строительство, реконструкцию, которые приводят к увеличению их первоначальной стоимости, приобретение машин,

оборудования, транспортных средств, производственного и хозяйственного инвентаря, инвестиции в объекты интеллектуальной собственности.

Для реализации инвестиций привлекается рабочая сила, количество и качество которой определяют такие факторы, как «Уровень занятости населения» и «Индекс производительности труда». Эффективность инвестиций зависит также от социально-экономического состояния страны, которое можно охарактеризовать остальными признаками: ВРП, оборот средних организаций, численность действующих ИП, число реализованных турпакетов населению, оборот розничной, объем ликвидированных основных фондов за год.

**Цель исследования**: произвести статистический анализ и построить многомерную эконометрическую модель инвестиций в основной капитал от выбранных факторов.

Для свертки данных по времени я использовал среднюю арифметическую и в некоторых случаях темп прироста, так как распределение не соответствовало нормальному.

**Инвестиции в основной капитал (табл. 1):** коэффициент вариации высокий – 93%, среднее значение отклонено от медианы, коэффициенты эксцесса и асимметрии высокие, следовательно, распределение близко к нормальному, но более островершинное и смещено вправо.

Таблица 1 – Описательная статистика и гистограмма среднего темпа роста инвестиций в основной капитал

|  |  |
| --- | --- |
| Средняя | 73685,41 |
| Медиана | 54678,44 |
| Эксцесс | 9,123646 |
| Асимметрия | 2,9656 |
| Размах | 347049,2 |
| Минимум | 22039,98 |
| Максимум | 369089,2 |
| Коэффициент  вариации | 93,00% |



**Уровень занятости населения (табл. 2):** коэффициент вариации незначительный – 7,15%, среднее значение практически совпадает с медианой и модой, данные варьируются от 49,8% до 78,3%.

Таблица 2 – Описательная статистика и гистограмма уровня занятости населения

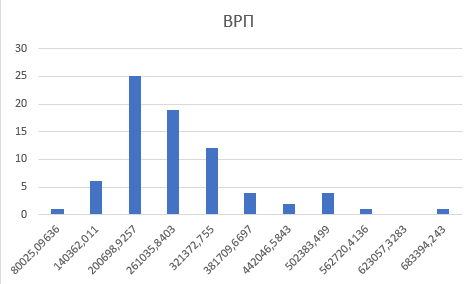
|  |  |
| --- | --- |
| Средняя | 63,58625 |
| Медиана | 63,58462 |
| Эксцесс | 1,600946 |
| Асимметрия | 0,130629 |
| Размах | 28,51538 |
| Минимум | 49,82308 |
| Максимум | 78,33846 |
| Коэффициент  вариации | 7,15% |



**ВРП (табл. 3):** коэффициент вариации высокий – 84,42%, среднее значение отклонено от медианы, экстремальное значение размаха.

Таблица 3 – Описательная статистика и гистограмма среднедушевых денежных доходов

|  |  |
| --- | --- |
| Средняя | 302983,4 |
| Медиана | 238499,9 |
| Эксцесс | 9,731802 |
| Асимметрия | 3,025237 |
| Размах | 1415164 |
| Минимум | 80025,1 |
| Максимум | 1495189 |
| Коэффициент  вариации | 84,42% |



**Среднедушевые денежные доходы (табл. 4):** коэффициент вариации высокий – 37,75%, среднее значение отклонено от медианы, данные варьируются от 9493 до 49093 рублей. Экстремальные значения принадлежат г. Москве и Ямало-Ненецкому АО.

Таблица 4 – Описательная статистика и гистограмма среднедушевых денежных доходов

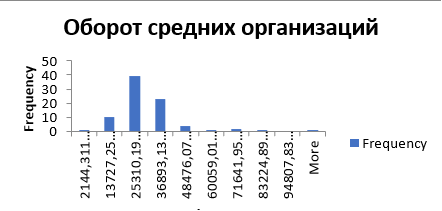
|  |  |
| --- | --- |
| Средняя | 17852,6 |
| Медиана | 15961,56 |
| Эксцесс | 6,81984 |
| Асимметрия | 2,349861 |
| Размах | 39600,44 |
| Минимум | 9492,866 |
| Максимум | 49093,31 |
| Коэффициент  вариации | 37,75% |



**Оборот средних организаций (табл. 5):** коэффициент вариации высокий – 60,3%, среднее значение близко к медиане, данные варьируются от 2144 до 106391 рублей. Экстремально высокое значение принадлежит республике Саха.

Таблица 5 – Описательная статистика и гистограмма оборота средних организаций

|  |  |
| --- | --- |
| Средняя | 26108,69 |
| Медиана | 23411,29 |
| Эксцесс | 9,610925 |
| Асимметрия | 2,506742 |
| Размах | 104246,5 |
| Минимум | 2144,311 |
| Максимум | 106390,8 |
| Коэффициент  вариации | 60,33% |



**Численность ИП (табл. 6):** коэффициент вариации средний – 18,03%, среднее значение близко к медиане, данные варьируются от 1094 до 3012.

Таблица 6 – Описательная статистика и гистограмма численности ИП

|  |  |
| --- | --- |
| Средняя | 1840,174 |
| Медиана | 1744,211 |
| Эксцесс | 1,769368 |
| Асимметрия | 1,028909 |
| Размах | 1918,127 |
| Минимум | 1094,398 |
| Максимум | 3012,525 |
| Коэффициент  вариации | 18,03% |



**Общее количество преступлений (табл. 7):** коэффициент вариации значительный – 30,94%, среднее значение близко к медиане, данные варьируются от 260 до 2880.

Таблица 7 – Описательная статистика и гистограмма общего количества преступлений

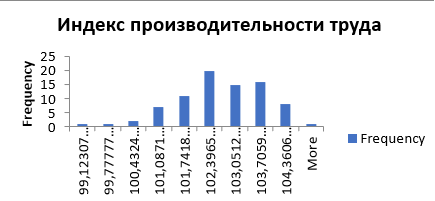
|  |  |
| --- | --- |
| Средняя | 1535,749 |
| Медиана | 1476,06 |
| Эксцесс | 0,123525 |
| Асимметрия | 0,114829 |
| Размах | 2620,514 |
| Минимум | 259,6844 |
| Максимум | 2880,198 |
| Коэффициент  вариации | 30,94% |



**Индекс производительности труда (табл. 8):** коэффициент вариации незначительный – 1,07%, среднее значение практически совпадает с медианой, данные сбалансированы.

Таблица 8 – Описательная статистика и гистограмма индекса производительности труда

|  |  |
| --- | --- |
| Средняя | 102,3742 |
| Медиана | 102,35 |
| Эксцесс | 0,00021 |
| Асимметрия | -0,20013 |
| Размах | 5,892308 |
| Минимум | 99,12308 |
| Максимум | 105,0154 |
| Коэффициент  вариации | 1,07% |



**Число реализованных турпакетов (табл. 9):** коэффициент вариации высокий – 97,08%, среднее значение отклонено от медианы, данные варьируются от 108 до 11304. Экстремальные значения принадлежат г. Москве, Санкт-Петербургу и республике Саха. Данные не сбалансированы.

|  |  |
| --- | --- |
| Средняя | 2172,593 |
| Медиана | 1630,052 |
| Эксцесс | 7,526436 |
| Асимметрия | 2,591172 |
| Размах | 11196,06 |
| Минимум | 107,9763 |
| Максимум | 11304,04 |
| Коэффициент  вариации | 97,08% |

Таблица 9 – Описательная статистика и гистограмма количества реализованных турпакетов



**Розничная торговля (табл. 10):** коэффициент вариации значительный – 27,18%, среднее значение близко к медиане, данные варьируются от 43527 до 235723 рублей.

Таблица 10 – Описательная статистика и гистограмма объема розничной торговли

|  |  |
| --- | --- |
| Средняя | 109929,1 |
| Медиана | 108459,3 |
| Эксцесс | 3,627935 |
| Асимметрия | 1,014881 |
| Размах | 192196,7 |
| Минимум | 43526,66 |
| Максимум | 235723,3 |
| Коэффициент  вариации | 27,18% |



**Ликвидация основных фондов (табл. 11):** коэффициент вариации высокий – 104,28%, среднее значение отклонено от медианы, данные варьируются от 1071 до 32161 рублей. Экстремальные значения принадлежат г. Москве, Тюменской области, Ханты-Мансийскому и Ямало-Ненецкому АО.

Таблица 11 – Описательная статистика и гистограмма ликвидации основных фондов

|  |  |
| --- | --- |
| Средняя | 4538,436 |
| Медиана | 3257,199 |
| Эксцесс | 17,84417 |
| Асимметрия | 3,935156 |
| Размах | 31090,6 |
| Минимум | 1070,663 |
| Максимум | 32161,27 |
| Коэффициент  вариации | 104,28% |



**1.2. Корреляционный анализ**

По полученным оценкам (таблица 12) видно, что инвестиции в основной фонд сильно коррелируют с ВРП (х2), ликвидацией основных фондов (х10) и среднедушевыми денежными доходами (х3).

Таблица 12 – Матрица парных коэффициентов корреляции

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | y | x1 | x2 | x3 | x4 | x5 | x6 | x7 | x8 | x9 | x10 |
| y | 1,00 | 0,55 | 0,95 | 0,77 | 0,53 | 0,10 | 0,07 | -0,24 | 0,18 | 0,53 | 0,87 |
| x1 | **0,55** | 1,00 | 0,62 | 0,74 | 0,59 | -0,01 | 0,00 | 0,11 | 0,33 | 0,59 | 0,52 |
| x2 | 0,95 | 0,62 | 1,00 | 0,83 | 0,54 | 0,07 | 0,06 | -0,28 | 0,23 | 0,63 | 0,91 |
| x3 | 0,77 | 0,74 | 0,83 | 1,00 | 0,67 | 0,03 | 0,05 | -0,16 | 0,39 | 0,74 | 0,76 |
| x4 | 0,53 | 0,59 | 0,54 | 0,67 | 1,00 | 0,19 | 0,03 | -0,02 | 0,39 | 0,38 | 0,43 |
| x5 | 0,10 | -0,01 | **0,07** | 0,03 | 0,19 | 1,00 | 0,06 | 0,10 | 0,00 | -0,04 | -0,06 |
| x6 | 0,07 | **0,00** | 0,06 | 0,05 | **0,03** | 0,06 | 1,00 | -0,09 | 0,20 | -0,02 | 0,06 |
| x7 | **-0,24** | 0,11 | -0,28 | -0,16 | **-0,02** | 0,10 | -0,09 | 1,00 | -0,05 | 0,02 | -0,36 |
| x8 | **0,18** | 0,33 | **0,23** | **0,39** | 0,39 | **0,00** | 0,20 | -0,05 | 1,00 | 0,47 | 0,18 |
| x9 | **0,53** | **0,59** | 0,63 | 0,74 | **0,38** | -0,04 | -0,02 | 0,02 | 0,47 | 1,00 | 0,49 |
| x10 | 0,87 | **0,52** | 0,91 | 0,76 | **0,43** | -0,06 | **0,06** | -0,36 | 0,18 | **0,49** | 1,00 |

Таблица 13 – Матрица частных коэффициентов корреляции

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | y | x1 | x2 | x3 | x4 | x5 | x6 | x7 | x8 | x9 | x10 |
| y | - | -0,20 | 0,74 | 0,05 | 0,10 | 0,01 | 0,02 | 0,23 | -0,02 | -0,22 | 0,01 |
| x1 | -0,20 | - | 0,24 | 0,32 | 0,17 | -0,14 | -0,03 | 0,37 | 0,04 | -0,04 | -0,06 |
| x2 | 0,74 | 0,24 | - | 0,02 | 0,05 | 0,20 | 0,04 | -0,26 | -0,13 | 0,41 | 0,53 |
| x3 | 0,05 | 0,32 | 0,02 | - | 0,44 | 0,02 | 0,07 | -0,13 | -0,02 | 0,49 | 0,25 |
| x4 | 0,10 | 0,17 | 0,05 | 0,44 | - | 0,14 | -0,11 | 0,05 | 0,30 | -0,35 | -0,21 |
| x5 | 0,01 | -0,14 | 0,20 | 0,02 | 0,14 | - | 0,04 | 0,12 | 0,01 | -0,14 | -0,29 |
| x6 | 0,02 | -0,03 | 0,04 | 0,07 | -0,11 | 0,04 | - | -0,02 | 0,26 | -0,17 | -0,05 |
| x7 | 0,23 | 0,37 | -0,26 | -0,13 | 0,05 | 0,12 | -0,02 | - | -0,14 | 0,26 | -0,09 |
| x8 | -0,02 | 0,04 | -0,13 | -0,02 | 0,30 | 0,01 | 0,26 | -0,14 | - | 0,40 | 0,08 |
| x9 | -0,22 | -0,04 | 0,41 | 0,49 | -0,35 | -0,14 | -0,17 | 0,26 | 0,40 | - | -0,33 |
| x10 | 0,01 | -0,06 | 0,53 | 0,25 | -0,21 | -0,29 | -0,05 | -0,09 | 0,08 | -0,33 | - |

В таблице 13 зеленым цветом обозначены значимые коэффициенты (доверительный интервал которых не содержит ноль). Наибольшее количество значимых коэффициентов принадлежит ВРП (х2) и инвестициям в основной фонд (у). Также в таблице 12 оранжевым цветом выделены пары показателей, чьи парные и частные парные коэффициенты корреляции имеют разный знак.

Таблица 14 – Множественные коэффициенты корреляции

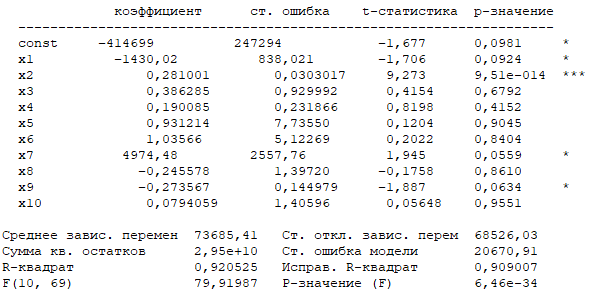
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | R | R^2 | F |  | R | R^2 | F |
| Инвестиции в основной капитал | 0,96 | 0,92 | 82,24 | Преступность | 0,30 | 0,09 | 0,68 |
| Уровень занятости населения | 0,80 | 0,64 | 12,78 | Индекс производительности | 0,59 | 0,34 | 3,74 |
| ВРП | 0,98 | 0,96 | 170,05 | Число реализованных турпакетов | 0,61 | 0,38 | 4,30 |
| Среднедушевые денежные доходы | 0,93 | 0,86 | 44,60 | Розничная торговля | 0,85 | 0,73 | 18,79 |
| Оборот средних организаций | 0,77 | 0,60 | 10,58 | Ликвидация основных фондов | 0,94 | 0,88 | 51,02 |
| Численность ИП | 0,42 | 0,18 | 1,55 | Fтаб = 1,97 |  |  |  |

Наиболее высокие коэффициенты детерминации принадлежат инвестициям в основной фонд и ВРП (таблица 14), можно сделать вывод, что остальные факторы объясняют инвестиции в основной фонд на 96%.

**1.3. Построение регрессионной модели.**

Построенная регрессия инвестиций в основной фонд (изображение 1) от всех факторов показала коэффициент детерминации, равный 92 %. Многие параметры статистически незначимы, а коэффициент отрицательный, что мешает интерпретации полученной модели. Скорректированный коэффициент детерминации отличается на 0,15%. Стандартная ошибка

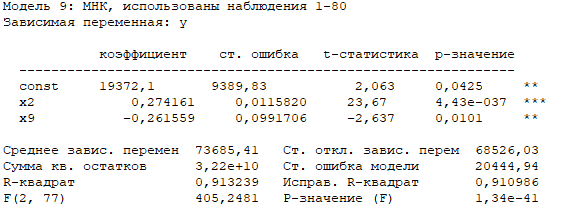
Изображение 1 – Регрессия инвестиций в основной фонд от всех факторов



модели составляет 20670 рублей, что достаточно много (медиана данного показателя равна 54678).

Проведем метод a-pasteriori на уровне значимости 5%.

Изображение 2 – Регрессия инвестиций в основной фонд после отбора факторов



Штраф за количество факторов уменьшился до 0,12%, все параметры значимы, коэффициент детерминации незначительно уменьшился- 91,3%, Константа значима и положительна (изображение 2). Исходя из полученной модели, инвестиции в основной фонд зависят от ВРП и объема розничной торговли, при нулевых значениях данных показателей инвестиции в основной фонд составит 19372 рубля на душу населения. При увеличении ВРП инвестиции в основной капитал растут, а при увеличении объема розничной торговли – уменьшаются, что противоречит экономическому смыслу.

Диаграмма рассеивания инвестиций в основной фонд и уровня занятости, розничной торговли, ликвидации основных фондов (диаграммы 1-3) не показывают явной линейной зависимости, необходимо проверить качество модели при изменении функционала. Проведем статистический анализ инвестиций в основной капитал после логарифмического преобразования.

Диаграммы 1-3 – Точечная диаграмма инвестиций в основной фонд от 1-уровня занятости, 2-розничной торговли, 3-ликвидации основных фондов

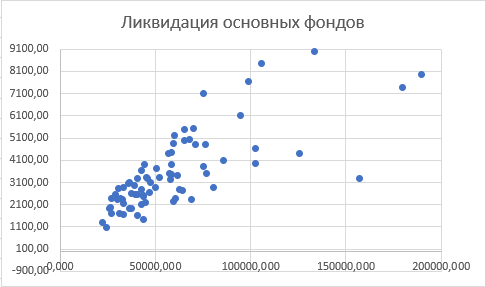
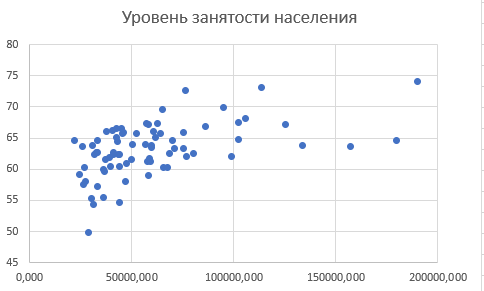
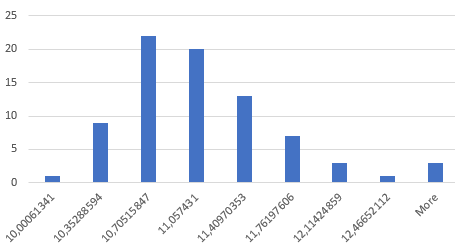


Таблица 15 – Описательная статистика и гистограмма логарифма от инвестиций в основной капитал

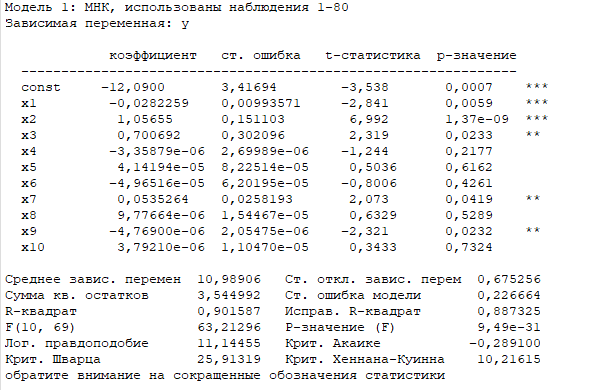
|  |  |
| --- | --- |
| Средняя | 10,98 |
| Медиана | 10,91 |
| Эксцесс | 4,543 |
| Асимметрия | 1,737 |
| Размах | 3,955 |
| Минимум | 10 |
| Максимум | 13,96 |
| Коэффициент | 6,1% |



Коэффициент вариации значительно уменьшился. Коэффициенты эксцесса и асимметрии сократились в два раза, следовательно, распределение ближе к нормальному, чем было до логарифмирования.

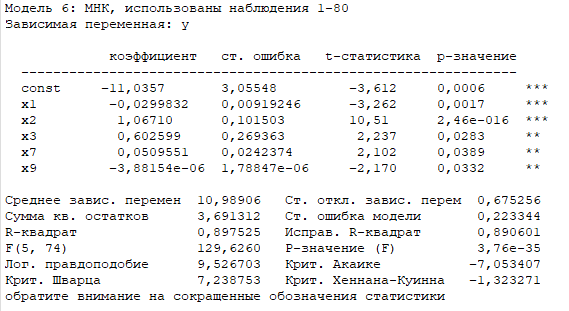
После изменения функционала понизился коэффициент детерминации до 90,2% (изображение 3), штраф составил 1,4%. Проведем метод a-pasteriori.

Изображение 3 – Регрессия инвестиций в основной фонд с измененным функционалом



Вследствие выбора другого функционала коэффициент детерминации понизился на 1.5% (изображение 4), изменился набор отобранных факторов – в модель были включены: х1 (уровень занятости населения), х3 (среднедушевые денежные доходы) и х7 (индекс производительности). Константа статистически значима, но отрицательна.

Изображение 4 – Регрессия логарифма инвестиций в основной фонд после отбора факторов



Исходя из коэффициентов регрессии можно объяснить зависимость между инвестициями в основной фонд и факторами. При росте ВРП, среднедушевых денежных доходов и производительности труда инвестиции увеличиваются, при увеличении уровня занятости населения и объема розничной торговли – уменьшаются, что противоречит экономическому смыслу.

**Глава 2. Эффект мультиколлинеарности.**

Для последующего построения эконометрической модели необходимо проверить исследуемые данные на мультиколлинеарность, так как данная проблема нарушает обратимость матрицы .

Таблица 16 – Матрица парных корреляций объясняющих переменных

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *x1* | *x2* | *x3* | *x4* | *x5* | *x6* | *x7* | *x8* | *x9* | *x10* |
| x1 | 1,00 | 0,62 | 0,74 | 0,59 | -0,01 | 0,00 | 0,11 | 0,33 | 0,59 | 0,52 |
| x2 | 0,62 | 1,00 | 0,83 | 0,54 | 0,07 | 0,06 | -0,28 | 0,23 | 0,63 | 0,91 |
| x3 | 0,74 | 0,83 | 1,00 | 0,67 | 0,03 | 0,05 | -0,16 | 0,39 | 0,74 | 0,76 |
| x4 | 0,59 | 0,54 | 0,67 | 1,00 | 0,19 | 0,03 | -0,02 | 0,39 | 0,38 | 0,43 |
| x5 | -0,01 | 0,07 | 0,03 | 0,19 | 1,00 | 0,06 | 0,10 | 0,00 | -0,04 | -0,06 |
| x6 | 0,00 | 0,06 | 0,05 | 0,03 | 0,06 | 1,00 | -0,09 | 0,20 | -0,02 | 0,06 |
| x7 | 0,11 | -0,28 | -0,16 | -0,02 | 0,10 | -0,09 | 1,00 | -0,05 | 0,02 | -0,36 |
| x8 | 0,33 | 0,23 | 0,39 | 0,39 | 0,00 | 0,20 | -0,05 | 1,00 | 0,47 | 0,18 |
| x9 | 0,59 | 0,63 | 0,74 | 0,38 | -0,04 | -0,02 | 0,02 | 0,47 | 1,00 | 0,49 |
| x10 | 0,52 | 0,91 | 0,76 | 0,43 | -0,06 | 0,06 | -0,36 | 0,18 | 0,49 | 1,00 |

По таблице 16 видно, что парные коэффициенты корреляции между некоторыми факторами (x2-x10, x2-x10, x3-x10) достаточно высокие, что говорит о мультиколлинеарности. Определитель матрицы равен 1,51E-82, то есть близок к нулю, что также подтверждает наличие мультиколлинеарности.

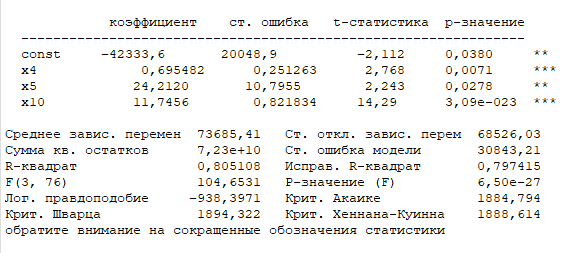
Таблица 17 – Множественные частные коэффициенты корреляции факторов, F-статистика и VIF критерий

|  |  |  |  |
| --- | --- | --- | --- |
|  | R^2 | F | VIF |
| x1 | 0,63 | 11,63 | 2,69 |
| x2 | 0,91 | 69,74 | 11,11 |
| x3 | 0,86 | 43,22 | 7,26 |
| x4 | 0,59 | 10,12 | 2,47 |
| x5 | 0,18 | 1,50 | 1,22 |
| x6 | 0,09 | 0,66 | 1,10 |
| x7 | 0,31 | 3,08 | 1,45 |
| x8 | 0,38 | 4,18 | 1,61 |
| x9 | 0,71 | 17,03 | 3,47 |
| x10 | 0,88 | 49,58 | 8,19 |

(1)

Статистика χ2 (рассчитанная по формуле 1) равна 90,75 – превышает критическое значение, равное 61,66. В таблице 17 представлены рассчитанные для объясняющих переменных F-критерии, восемь из которых превышают критическое значение (на уровне значимости 5% приблизительно 1,97), также присутствуют факторы, показатели VIF которых отличаются от единицы, VIF критерий для фактора х2 (ВРП) превышает критическое значение (больше 10), что в очередной раз показывает наличие проблемы мультиколлинеарности.

Изображение 5 – Модель МНК с переменными, отобранными по VIF-критерию



Попытаемся избавиться от мультиколлинеарности путем исключения переменных с высоким показателем VIF – х2 (ВРП) и х10 (ликвидация основных фондов). В результате была получена следующая модель (изображение 5).

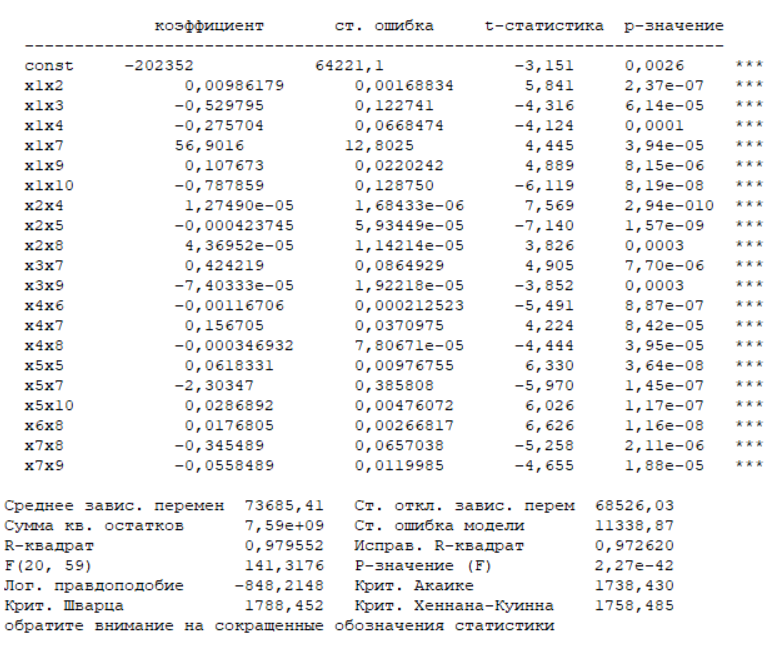
Наблюдается высокий R-квадрат, но ниже предыдущих моделей, статистически значимые коэффициенты регрессии, константа значима, однако отрицательная, эффект мультиколлинеарности был устранен.

В данном случае проблема мультиколлинеарности также может быть решена методом построения мультипликативной модели. С помощью кода на Python реализуем данный метод.

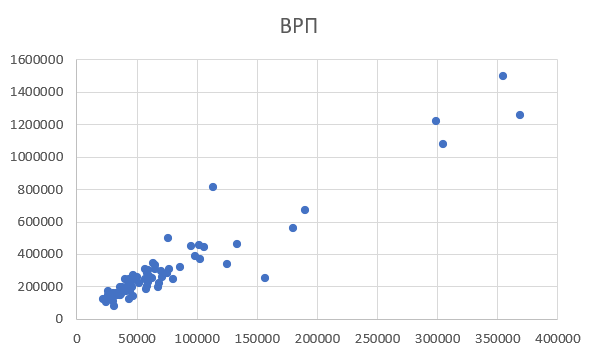
В результате была получена модель (изображение 6).

R-квадрат высокий, но эффект мультиколлинеарности остался и, так как количество факторов довольно большое и они представляют собой комбинации исходных, данную модель практически невозможно интерпретировать. Следовательно, несмотря на высокий коэффициент R2 и значимость коэффициентов регрессии, данную модель будет сложно использовать для анализа данных.

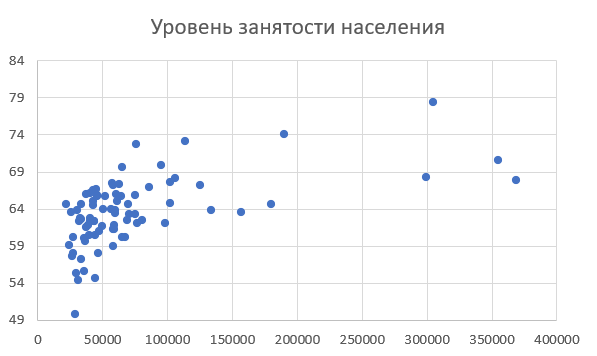
Изображение 6 – Мультипликативная модель



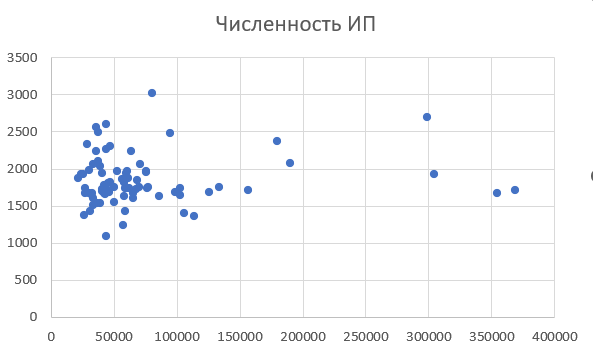
**Глава 3. Робастное статистическое оценивание**

*Диаграммы 4-5 – Инвестиции в основной фонд; ВРП* 

*Диаграммы 6-7 – Уровень занятости населения; среднедушевые денежные доходы*

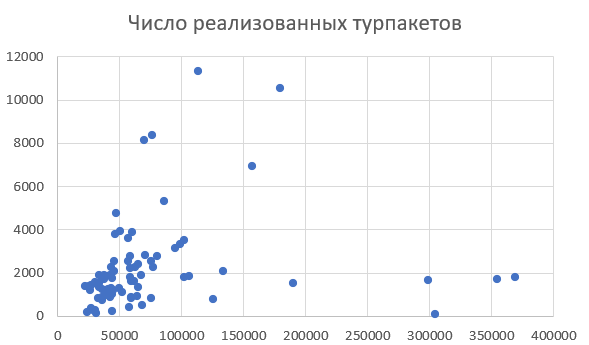
*Диаграммы 7-8 – оборот средних организаций; Численность ИП*

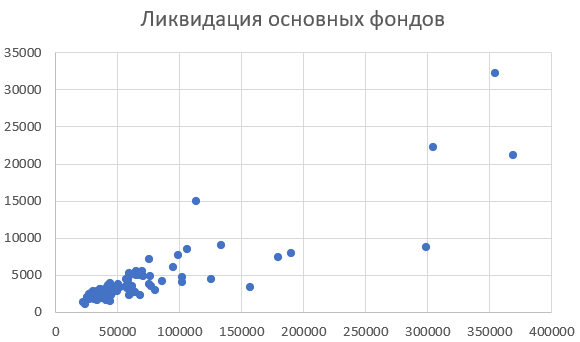
*Диаграммы 9-10 – Общее количество преступлений; индекс производительности*

*Диаграммы 10-11– Число реализованных турпакетов; розничная торговля*

*Диаграмма 12 – ликвидация основных фондов*

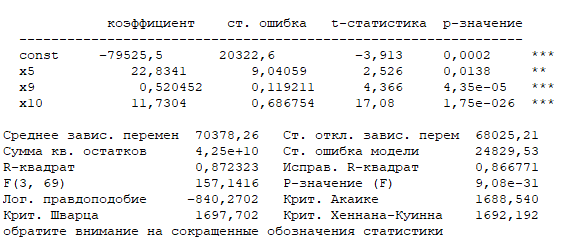


По визуальным данным (диаграммы 4-12) аномально высокие значения принадлежат Тюменской и Сахалинской областям, Ханты-Мансийскому и Чукотскому АО, а также г. Москве и республике Саха и некоторые другие города.

Нетипичными данными обладают следующие регионы: г. Москва, Тюменская область, Ханты-Мансийский АО, республика Саха, Амурская область, Сахалинская область и Чукотский АО. Также по одному нетипичному значению содержат г. Санкт-Петербург, Краснодарский край, Астраханская область, республика Дагестан, Чеченская республика, Томская область, Забайкальский край, Приморский край, Магаданская область.

Исключая различные комбинации регионов с подозреваемыми выбросами, построим модели и выберем модель с наивысшим коэффициентом R2:

Изображение 7 – Робастная регрессия без г. Москвы, Амурской области и Краснодарского края



В данной модели были исключены: г. Москва, Краснодарский край и Амурская область. R-квадрат довольно высокий – 87,3%, коэффициенты значимы, экономический смысл модели не нарушается. Однако признаки х10 и х3 сильно коррелируют между собой.. Попытаемся построить модель без коррелированных признаков и не менее высоким коэффициентом R2:

В данной модели (изображение 8) не были учтены г. Москва, г. Санкт-Петербурга и Амурская область. По сравнению с предыдущей моделью константа является статистически значимой, но количество факторов сильно сократилось с 5 до 1. Разница в коэффициентах R-квадрат небольшая – менее 1%. Константа отрицательная, по данной модели можно только сказать, что при росте ВРП увеличиваются инвестиции в основной фонд, данная закономерность прослеживается во всех моделях. Модель построена на одном факторе, следовательно, от данной модели придется отказаться.

(2)

*– среднее арифметическое, s – выборочное среднеквадратичное отклонение*

Рассчитаем T-статистику Смирнова-Граббса по формуле 2 и выделим наблюдения, которые не прошли тест (превысили критическое значение 3,160 для 2,5% уровня значимости):

Изображение 8 – Регрессионная модель без учета г.Москвы и Санкт- Петербурга, Амурской области

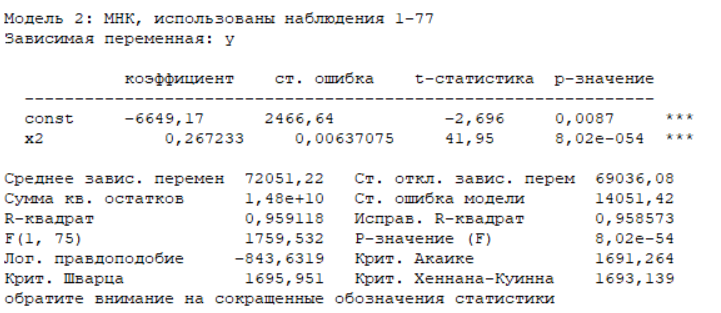
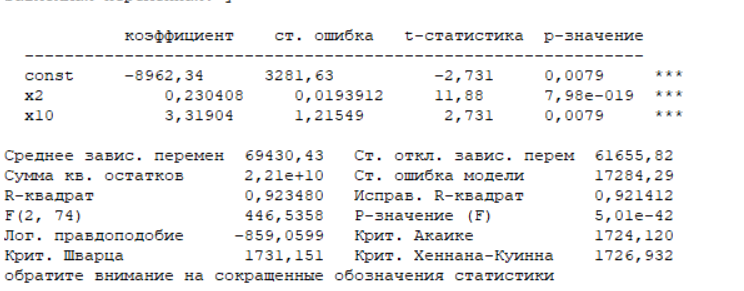


Таблица 18 – Статистика Смирнова-Граббса

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | y | x1 | x2 | x3 | x4 | x5 | x6 | x7 | x8 | x9 | x10 |
| г, Москва | 0,58 | 2,09 | 1,98 | 3,53 | 0,56 | -1,44 | -0,43 | -1,28 | 4,33 | 4,21 | 2,20 |
| Краснодарский край | 0,10 | -0,25 | -0,23 | 0,29 | 0,03 | 3,53 | -0,55 | -0,82 | 0,29 | 1,04 | -0,35 |
| Тюменская область | 4,31 | 0,93 | 3,74 | 1,70 | 0,39 | -0,39 | -1,54 | -0,83 | -0,19 | 1,69 | 3,52 |
| Ханты-Мансийский АО | 4,10 | 1,53 | 4,66 | 2,07 | 0,49 | -0,49 | -0,08 | -2,97 | -0,21 | 1,69 | 5,84 |
| Республика Саха | 1,55 | 0,24 | 1,00 | 1,12 | 5,10 | 1,61 | -0,74 | 0,09 | 3,96 | 0,38 | 0,59 |
| Сахалинская область | 3,29 | 1,02 | 3,58 | 2,24 | 1,53 | 2,56 | 1,27 | -0,75 | -0,25 | 2,71 | 0,88 |
| Чукотский АО | 3,37 | 3,24 | 3,02 | 4,64 | 3,47 | 0,26 | 0,04 | -1,30 | -0,98 | 0,17 | 3,73 |

Изображение 8 – Регрессионная модель по типичным данным, отобранным статистикой Смирнова-Граббса



Построим модели, последовательно исключая по одному подозреваемому в выбросах региону и выберем лучшую модель (изображение 9)

После исключения из модели Краснодарского края, г. Москвы и Ханты-мансийского АО R-квадрат стал равен 92,3%, при этом все коэффициенты регрессионной модели уверенно значимы на уровне 1%. Экономический смысл данной модели заключается в том, что при росте ВРП и ликвидаций основных фондов растут инвестиции в основной фонд, противоречий нет. Однако признаки х2 и х10 довольно сильно коррелируют между собой (таблица 16).

Также проверим данные на выбросы критерием Титьена-Мура. Для этого необходимо рассчитать соответствующие статистики по формуле 3:

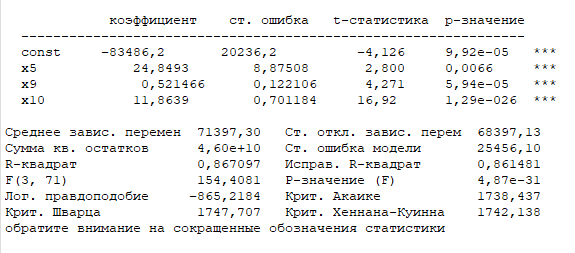
, – усеченное выборочное среднее (3)

По данному критерию были определены следующие совокупности наблюдений как выбросы:

* По у (инвестиции в основной капитал) – Сахалинская и Тюменская области, Чукотский и Ханты-Мансийский АО.
* По х2 (ВРП) – Тюменская и Воронежская области, Приморский край и Чукотский АО
* По х3 (Среднедушевые денежные доходы) – Тюменская, Рязанская и Сахалинская области, республика Коми, Приморский край и Чукотский АО.
* По х4 (оборот средних организаций) – г. Москва, Тюменская и Сахалинская области, республика Татарстан и Чукотский АО.
* По х8 (Число реализованных турпакетов) – Амурская, Омская, Рязанская и Магаданская области и г. Москва.
* По х10 (Ликвидация основных фондов) – Воронежская и Тюменская области и Приморский край

Также последовательно исключая из набора данных выделенные выше наблюдения, построим модели и определим лучшую из них:

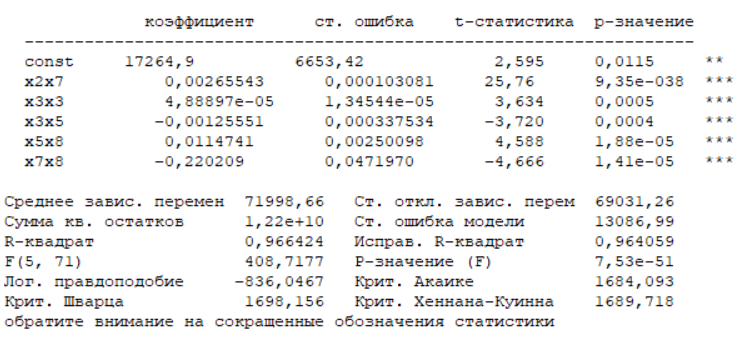
Изображение 9 – Регрессионная модель по типичным данным, отобранным статистикой Титьена-Мура



Было исключено 5 регионов (изображение 9): г. Москва, Магаданская, Амурская, Омская и Рязанская области. R-квадрат довольно высокий, константа значима, но отрицательна, мультиколлинеарности нет, коэффициенты значимы, экономический смысл не нарушен.

Для неочищенных данных метод мультипликативной модели показал неудовлетворительные результаты, попробуем применить данный метод для типичных данных.

Изображение 10 – Мультипликативная модель по типичным данным



Полученная модель (изображение 10) обладает довольно высоким коэффициентом детерминации, все коэффициенты значимы и константа положительная. Однако некоторые мультипликативные переменные остались коррелированными между собой (таблица 19).

Таблица 19 – Матрица парных коэффициентов корреляции для мультипликативных признаков

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | ***x2x7*** | ***x3x3*** | ***x3x5*** | ***x5x8*** | ***x7x8*** |
| **x2x7** | 1 |  |  |  |  |
| **x3x3** | 0,804693 | 1 |  |  |  |
| **x3x5** | 0,782359 | 0,895196 | 1 |  |  |
| **x5x8** | 0,171199 | 0,153016 | 0,321755 | 1 |  |
| **x7x8** | 0,156387 | 0,137399 | 0,250067 | 0,975378 | 1 |

* х2х7 – произведение ВРП (на душу населения) и индекса производительности труда
* х3х3 – квадрат среднедушевых денежных доходов (на душу населения)
* х3х5 – произведение среднедушевых денежных доходов и численности ИП (на 100 тыс. человек)
* х5х8 – произведение численности ИП (на 100 тыс. человек) и реализованных турпакетов (на 100 тыс. человек)
* х7х8 – произведение индекса производительности труда и численности реализованных турпакетов (на 100 тыс. человек)

Несмотря на небольшое количество факторов и положительную константу, могут возникнуть проблемы с интерпретацией модели. Так, например, произведение численности ИП и реализованных турпакетов лишено экономического смысла.

**Глава 4. Метод главных компонент.**

**4.1. Робастное оценивание с помощью метода главных компонент.**

Для выявления нетипичных значений также можно воспользоваться методом главных компонент. Данный метод нацелен на сокращении признакового пространства, за счет чего появляется возможность более точно определить выбросы графическим способом.

МГК будет осуществляться на корреляционной матрице (таблица 12). Для нахождения собственных значений и собственных векторов воспользуемся программным пакетом gretl.

Таблица 20 – Таблица Собственных значений и их долей дисперсии по всем данным

|  |  |  |  |
| --- | --- | --- | --- |
| **Компонента** | **Собственное значение** | **Доля** | **Интегральная** |
| 1 | 4,8975 | 0,4452 | 0,4452 |
| 2 | 1,5447 | 0,1404 | 0,5857 |
| 3 | 1,2204 | 0,1109 | 0,6966 |
| 4 | 1,072 | 0,0975 | 0,7941 |
| 5 | 0,8794 | 0,0799 | 0,874 |
| 6 | 0,5048 | 0,0459 | 0,9199 |
| 7 | 0,4062 | 0,0369 | 0,9568 |
| 8 | 0,295 | 0,0268 | 0,9836 |
| 9 | 0,1137 | 0,0103 | 0,994 |
| 10 | 0,0566 | 0,0051 | 0,9991 |
| 11 | 0,0097 | 0,0009 | 1 |

*Диаграмма 13 – График собственных значений главных компонент*



Первые три компоненты забирают 69,66% от суммарной дисперсии (таблица 20), соответственно в остальных компонентах содержится около 30% от суммарной дисперсии.

Изменение собственных значений довольно резкое между первыми двумя компонентами, последующие изменения равномерные (диаграмма 13). Доля первой компоненты составляет 44,5%, то есть практически половина от общей дисперсии, следующие четыре компоненты между собой слабо отличаются.

Таблица 21 – Матрица весов главных компонент

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | V1 | V2 | V3 | V4 | V5 | V6 | V7 | V8 | V9 | V10 | V11 |
| y | -0,379 | 0,339 | 0,046 | 0,202 | 0,054 | -0,086 | -0,291 | 0,056 | -0,451 | -0,314 | -0,546 |
| x1 | -0,344 | -0,218 | 0,21 | -0,154 | 0,125 | -0,094 | 0,556 | 0,649 | -0,069 | -0,065 | -0,045 |
| x2 | -0,412 | 0,284 | -0,003 | 0,153 | -0,004 | -0,039 | -0,129 | 0,045 | -0,151 | -0,156 | 0,812 |
| x3 | -0,437 | -0,02 | -0,009 | -0,028 | -0,018 | 0,11 | 0,031 | -0,193 | -0,226 | 0,837 | -0,07 |
| x4 | -0,249 | -0,462 | 0,051 | 0,08 | -0,265 | -0,718 | -0,008 | -0,337 | 0,059 | -0,111 | -0,009 |
| x5 | 0,013 | -0,254 | 0,19 | 0,793 | -0,361 | 0,345 | 0,074 | 0,109 | 0,049 | 0,01 | -0,022 |
| x6 | -0,029 | -0,162 | -0,525 | 0,416 | 0,684 | -0,076 | 0,194 | -0,111 | 0,009 | -0,017 | -0,003 |
| x7 | -0,014 | -0,141 | 0,729 | 0,062 | 0,556 | 0,006 | -0,3 | -0,099 | 0,177 | 0,049 | 0,037 |
| x8 | -0,177 | -0,564 | -0,297 | -0,199 | 0 | 0,248 | -0,601 | 0,311 | -0,013 | -0,055 | 0,029 |
| x9 | -0,36 | -0,166 | 0,041 | -0,236 | 0 | 0,516 | 0,291 | -0,527 | 0,051 | -0,389 | -0,045 |
| x10 | -0,392 | 0,296 | -0,134 | 0,059 | -0,052 | -0,029 | -0,097 | 0,118 | 0,824 | 0,044 | -0,176 |

Исходя из весов главных компонент (таблица 21) , их можно интерпретировать следующим образом:

* Первая главная компонента: приблизительно в равном соотношении содержит в себе информацию об инвестициях в основной капитал (у), уровне занятости населения (х1), ВРП (х2), среднедушевых денежных доходов (х3), розничной торговли (х9) и ликвидации основных фондов (х10). Наибольший вес принадлежит переменной х3 и х2 (среднедушевым денежным доходам и ВРП). В меньшей степени учитываются остальные показатели – оборот средних организаций, численность ИП, общее количество преступлений, индекс производительности труда и количество реализованных турпакетов.
* Вторая главная компонента преимущественно забирает информацию с переменных х8 и х4 (число реализованных турпакетов и оборот средних организаций), также в меньшей степени данная компонента содержит некоторые остальные показатели.
* Третья главная компонента отвечает за индекс производительности труда, а также за общее количество преступлений.

Так как веса первых двух главных компонент распределены довольно равномерно, раскрыть их смысл сложно, с помощью полученных главных компонент целесообразно будет лишь выделить показатели, которые несут в себе меньше информации, чем остальные – исходя из того, что первая компонента забирает практически половину суммарной дисперсии и что веса при переменных х5-х8 малы, можно сделать вывод о недостаточном количестве информации, содержащейся в соответствующих показателях (, численность ИП, общее количество преступлений индекс производительности труда и количество реализованных турпакетов).

Проведем визуальный анализ структуры данных и выявим нетипичные наблюдения графическим способом.

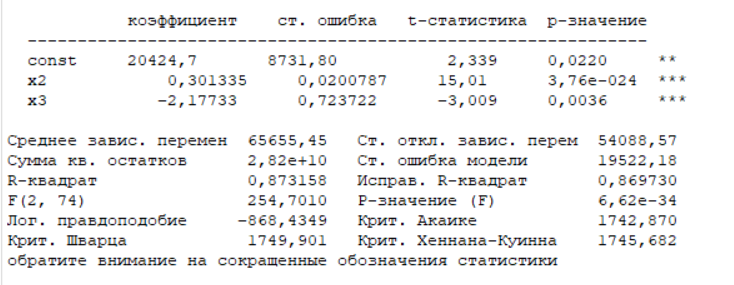
*Диаграммы 14-15 –* *Диаграмма рассеивания между первой и второй компонентой и диаграмма рассеивания между первой и третьей компонентой.*

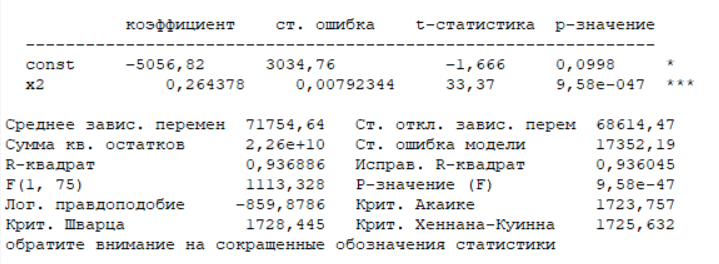
*Диаграмма 16 - Диаграмма рассеивания между второй и третьей компонентой.*



Изображение 11 – Регрессионная модель по типичным данным

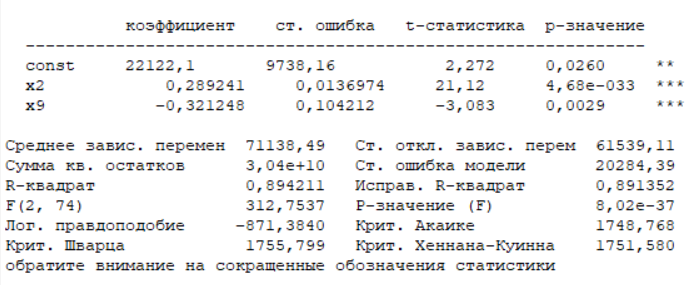


Изображение 13 – Регрессионная модель после исключения Приморского края, г. Москвы, г. Санкт-Петербурга и республики Саха



Полученные модели обладают незначимыми константами (изображения 12-13), за исключением последней модели (изображение 14) - при достаточно высоком R-квадрате обладает статистически значимыми коэффициентами, но отрицательные коэффициент при переменной, отвечающей за объем розничной торговли, нарушает экономический смысл, от данных моделей стоит отказаться.

Изображение 14 – Регрессионная модель после исключения Ханты-Мансийского АО, республики Тыва и Бурятии



**4.2. Применение метода главных компонент на факторных признаках.**

Построим главные компоненты, не включая объясняемую переменную у (инвестиции в основной фонд). Таким образом будет сокращено признаковое пространство и решена проблема мультиколлинеарности.

Аналогично предыдущим действиям воспользуемся программным пакетом gtretl для расчета собственных значений и векторов главных компонент, используя матрицу парных корреляций:

*Диаграмма 16 - График доли главных компонент факторных признаков*



Таблица 22 – Собственные значения главных компонент на факторных признаках

|  |  |  |  |
| --- | --- | --- | --- |
| **Компонента** | **Собственное значение** | **Доля** | **Интегральная** |
| 1 | 4,4123 | 0,4412 | 0,4412 |
| 2 | 1,3326 | 0,1333 | 0,5745 |
| 3 | 1,1398 | 0,114 | 0,6885 |
| 4 | 1,0473 | 0,1047 | 0,7932 |
| 5 | 0,7381 | 0,0738 | 0,867 |
| 6 | 0,5773 | 0,0577 | 0,9247 |
| 7 | 0,3058 | 0,0306 | 0,9553 |
| 8 | 0,2848 | 0,0285 | 0,9838 |
| 9 | 0,1083 | 0,0108 | 0,9946 |
| 10 | 0,0537 | 0,0054 | 1 |

Так же, как и в предыдущем случае, довольно большая разница между первой и второй главной компонентой, следующие четыре обладают приблизительно равной долей суммарной дисперсии.

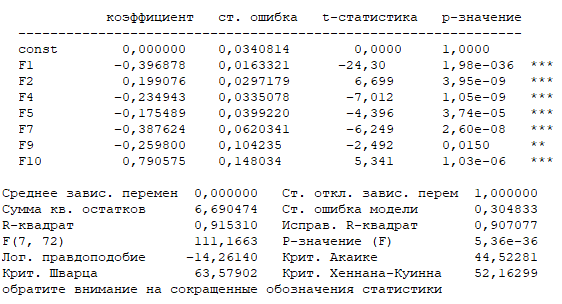
Таблица 23 – Матрица собственных векторов главных компонент на факторных признаках

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | V1 | V2 | V3 | V4 | V5 | V6 | V7 | V8 | V9 | V10 |
| x1 | -0,378 | -0,243 | 0,159 | 0,053 | -0,23 | -0,216 | 0,558 | 0,581 | -0,135 | -0,056 |
| x2 | -0,43 | 0,212 | 0,049 | -0,182 | -0,144 | 0,145 | -0,253 | 0,049 | -0,318 | 0,725 |
| x3 | -0,452 | -0,003 | 0,046 | -0,023 | -0,039 | 0,003 | 0,089 | -0,194 | 0,853 | 0,134 |
| x4 | -0,34 | -0,244 | -0,096 | -0,189 | 0,173 | -0,689 | -0,1 | -0,44 | -0,243 | -0,112 |
| x5 | -0,018 | -0,324 | -0,35 | -0,771 | 0,178 | 0,338 | 0,088 | 0,128 | 0,028 | -0,084 |
| x6 | -0,039 | 0,073 | -0,792 | 0,189 | -0,556 | -0,015 | 0,085 | -0,111 | -0,028 | -0,027 |
| x7 | 0,089 | -0,716 | 0,216 | 0,08 | -0,456 | 0,087 | -0,455 | -0,004 | 0,057 | 0,026 |
| x8 | -0,23 | -0,254 | -0,379 | 0,462 | 0,56 | 0,069 | -0,288 | 0,343 | 0,033 | 0,068 |
| x9 | -0,368 | -0,152 | 0,123 | 0,248 | 0,062 | 0,57 | 0,27 | -0,474 | -0,296 | -0,225 |
| x10 | -0,392 | 0,354 | 0,075 | -0,129 | -0,169 | 0,073 | -0,478 | 0,241 | -0,02 | -0,614 |

Веса собственного вектора первой компоненты (таблица 23) практически равны весам, полученным в предыдущих расчетах, соответственно соотношение содержащейся в первой компоненте информации идентичное, Вторая компонента в большей степени содержит информацию признака x7 (индекс производительности труда), а третья компонента – х6 (общее количество преступлений).

Построим регрессионную модель с наибольшим количеством значимых коэффициентов по всем регионам:

Изображение 15 – Регрессионная модель на главных компонентах



Так как модель имеет следующий вид:

(4)

– модельное стандартизированное значение, – матрица оцененных коэффициентов линейной регрессии, – матрица главных компонент.

С помощью собственных векторов соответствующих компонент выразим данную модель через исходные признаковые переменные X, в соответствии с уравнением:

, V - матрица собственных векторов (5)

(6)

Обозначим (7)

Таким образом с помощью вектора преобразованных коэффициентов станет возможным оценить интерпретируемость модели. Для модели на изображении 16 рассчитаем коэффициенты :

Таблица 24 – коэффициенты

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **b1** | **b2** | **b3** | **b4** | **b5** |
| -0,09594 | 1,034744 | 0,040869 | 0,113758 | -0,01525 |
| **b6** | **b7** | **b8** | **b9** | **b10** |
| 0,036159 | 0,065482 | -0,00928 | -0,15899 | -0,00892 |

Некоторые коэффициенты отрицательные (таблица 24) , то есть для признака х9 коэффициент регрессии равный -0,159 нарушает экономический смысл модели, так как при увеличении объема розничной торговли инвестиции также должны увеличиваться.

Методом исключения компонент из модели вернуть экономический смысл модели не получилось, далее применим МГК к типичным данным.

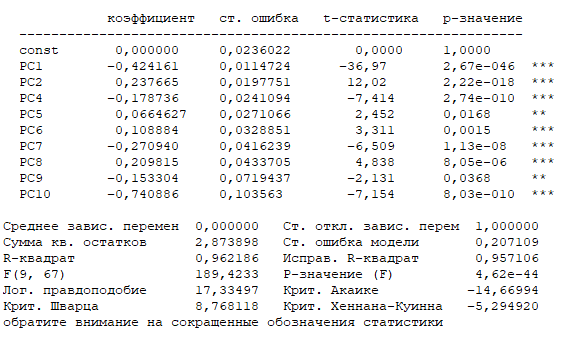
**4.3. Применение метода главных компонент на типичных данных.**

Также построим для сравнения модели для типичных данных с учетом робастных оценок с помощью МГК и робастных оценок, полученных в 3 главе.

Все построенные модели не превосходят по коэффициенту R-квадрат, при этом нарушается экономический смысл модели, знаки коэффициентов регрессии не совпадают с коэффициентами парной корреляции.

Например, модель, построенная с исключением наблюдений г. Москвы, Краснодарского края и Амурской области (изображение 16):

Изображение 16 – Регрессионная модель на главных компонентах



Рассчитаем коэффициенты по формуле 7:

Таблица 25– коэффициенты

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **b1** | **b2** | **b3** | **b4** | **b5** |
| -0,10266 | 0,971768 | 0,080715 | 0,048654 | -0,01212 |
| **b6** | **b7** | **b8** | **b9** | **b10** |
| -0,0074 | 0,051862 | -0,00159 | -0,07789 | 0,040661 |

Противоречивые коэффициенты при признаках b1 – уровень занятости населения, b5 – численность ИП, b9 – объем розничной торговли и другие.

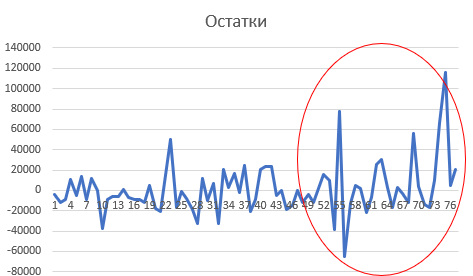
**Глава 5. Эффект гетероскедастичности остатков регрессионной модели.**

**5.1. Проверка модели на гетероскедастичность.**

В качестве лучшей модели была выбрана робастная модель (изображение 7), в которой были исключены наблюдения Краснодарского края, Амурской области и г. Москвы. Она не нарушает экономический смысл, обладает высоким коэффициентом R-квадрат, метод главных компонент давал в результате модель с коэффициентом детерминации, равным 96,22% (изображение 17), на 9% больше, чем у выбранной модели, однако модель, построенная на МГК, не подходит для прогнозирования, так как коэффициенты указывают на неверную зависимость. Далее будем работать с робастной моделью (изображение 7).

Перед расчетом прогнозных значений, необходимо убедиться в отсутствии проблемы гетероскедастичности. В первую очередь проанализируем остатки визуально:

*Диаграмма 17 - График остатков робастной модели*



Были выявлены скачки (диаграмма 17), что нарушает условие постоянства дисперсии ошибок. Осуществим проверку зависимости остатков от факторных переменных визуально:

*Диаграммы 18-19 – диаграмма рассеяния остатков от признаков х5, х9*

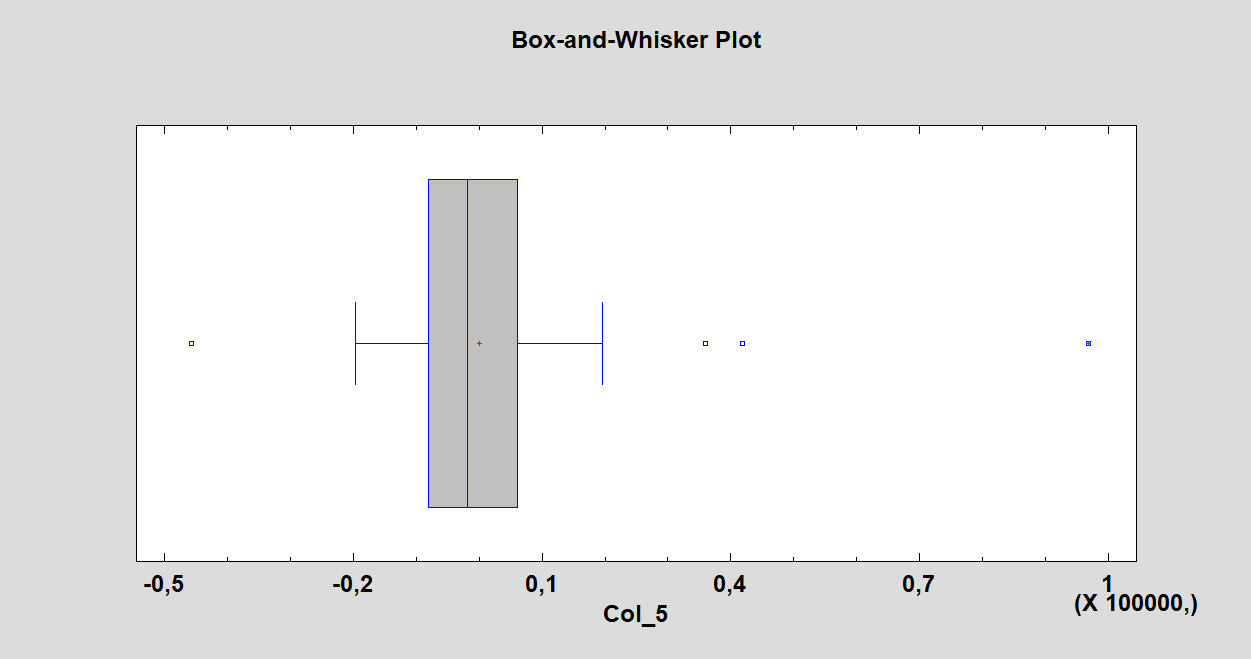
*Диаграмма 20 – диаграмма рассеяния остатков от признака х10*

По оси абсцисс расположены значения признаковых переменных, по оси ординат – остатки модели (диаграммы 18-20). На первый взгляд остатки зависят от факторов, проверим распределение остатков для проведения статистических тестов на выявление гетероскедастичности.

*Диаграмма 22 – гистограмма остатков*



*Диаграмма 21 – диаграмма «ящик с усами» остатков*



По гистограмме и диаграмме ящика с усами (диаграммы 21-22), можно сделать вывод, что данное распределение имеет более заостренную форму, а также присутствует асимметрия. Кроме того, ящик с усами показывает наличие выбросов, которые также были замечены ранее. Рассчитаем описательную статистику для остатков:

Таблица 26– Описательная статистика остатков

|  |  |  |  |
| --- | --- | --- | --- |
| **Среднее значение** | **Стандартное**  **отклонение** | **Коэффициент**  **вариации** | **Минимум** |
| 1479,57 | 25641,6 | 1733,05% | -64693,2 |
| **Максимум** | **Размах** | **Коэффициент ассиметрии** | **Эксцесс** |
| 96943 | 142635 | 8,86141 | 24,3183 |

По данным расчетам (таблица 26) можно сделать вывод, что остатки распределены не по нормальному закону. Соответственно корректной проверкой будет тест Спирмена, который не зависит от закона распределения остатков.

После присвоения рангов по остаткам и по признакам, рассчитаем сумму квадратов разницы присвоенных рангов для расчета ранговой корреляции Спирмена и т-статистику по следующим формулам 8-9:

(8-9)

В результате было получено значение -0,10177 для переменной х5, -0,16702 для переменной х9 и -0,22627 для переменной х10, критическое значение на доверительном уровне 5% для данной статистики равно t(0,05;n-2) = -1,9921. Следовательно, по тесту Спирмена остатки зависят от переменной х9, t-статистика по данному признаку равна -2,01177.

Также проверим остатки с помощью теста Уайта. Для этого построим модель следующего вида (формула 10):

(10)

Учитывая, что в модели задействуется 9 факторов, получим более 8 наблюдений на каждый, этого достаточно для корректной оценки. Проверим значимость данной модели по коэффициенту R-квадрат, который равен 0,907971153, то есть очевидно статистически значимый. Что также указывает на присутствие гетероскедастичности.

Далее проведем тест Голдфельда-Квандта - для этого нужно посчитать статистику Фишера по формуле

, (11) где - дисперсия остатков МНК по первым наблюдениям, - дисперсия остатков МНК по последним наблюдениям.

Для реализации данного теста разобьем на и равные 29, что составляет приблизительно от общей выборки. Далее считаем статистику Фишера последовательно сортируя по каждому фактору в порядке возрастания. В результате тест Голдфельда-Квандта определил гетероскедастичность по каждому фактору (таблица 27):

Таблица 27– Статистика Голдфельда-Квандта

|  |  |  |
| --- | --- | --- |
| F-x5 | F-x9 | F-x10 |
| 3,130692205 | 20,7694 | 13,49132 |

Несмотря на то, что остатки распределены не по нормальному закону, к чему тест Голдфельда-Квандта чувствителен, результаты теста совпадают с результатами предыдущих проверок, следовательно, остатки выбранной модели не обладают свойством гомоскедастичности.

**5.2. Устранение гетероскедастичности.**

Так как тест Уайта показал высокий коэффициент детерминации, воспользуемся квадратичным функционалом для расчета модельных квадратов остатков. После вычисления модельных значений квадратов остатков рассчитаем коэффициенты взвешенной модели наименьших квадратов:

, где (12)

В результате были получены следующие коэффициенты (таблица 28):

Таблица 28– коэффициенты ВМНК

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| 1,172231 | -4,11001 | 0,1556 | 9,963931 |

Посредством данной модели рассчитаем модельные значения , после чего извлечем из них модельный исходный результативный признак путем умножения на веса (корень из модельных остатков в квадрате).

После того, как были рассчитаны модельные значения исходного объясняемого признака, для сравнения была получена стандартная ошибка модели: RMSE = 200335,3, стандартная ошибка исходной модели составила 24829,5, что намного меньше, чем в ВМНК. От метода ВМНК придется отказаться.

Возможно тесты показали наличие гетероскедастичности из-за выбросов на графике остатков были заметны скачки, которые принадлежат Тюменской и Сахалинской областям.

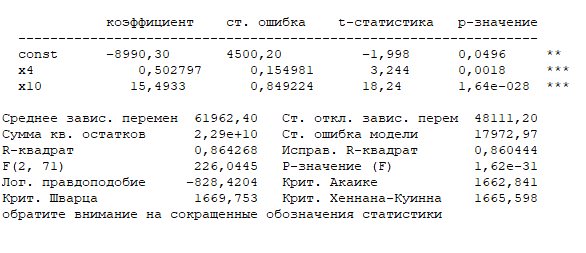
Признак х9 по проведенным тестам вызывает наибольшие подозрения, в очередной раз визуально проверим зависимость с объясняемым показателем. Как было ранее замечено, функционал не совсем похож на линейный (диаграмма 23), напоминает квадратичную зависимость – попробуем перестроить модель с учетом функционала.

*Диаграмма 23 – диаграмма рассеяния инвестиций в основной фонд от объема розничной торговли*

После изменения функционала проблема гетероскедастичности остатков осталась, следовательно, проблема в выбросах – Тюменская, Сахалинская, Магаданская области, Ханты-Мансийский и Чукотский АО и республика Саха.

Последовательно исключая выбросы была получена следующая модель (изображение 16):

Изображение 16 – Робастная регрессионная модель после исключения Сахалинской области, Чукотского и Ханты-Мансийского АО



По графику остатков можно предположить, что дисперсия остатков постоянная, после проверки остатков тестом Спирмена, были получены ранговые корреляции -0,041932 и -0,1941, которые статистически незначимы, следовательно, была принята гипотеза о гомоскедастичности остатков - необходимости в применении ВМНК нет (диаграмма 23).

*Диаграмма 23 – Остатки после избавления от выбросов*

Также рассчитаем статистику Дарбина-Уотсона по формуле 13:

(13)

DW = 2,044077, следовательно, автокорреляции нет. Так как в свертке, по которой была построена модель отсутствует измерение времени, данный результат ожидаемый, измерение времени было устранено, на моменте свертывания данных, когда были заменены динамические последовательности на среднее значение по всему промежутку времени. Тест Дарбина-Уотсона также подтвердил, что мы избавились от проблемы гетероскедастичности. Модель готова для прогнозирования.

**Глава 6. Расчет прогнозных значений.**

В результате всех преобразований была получена полностью статистически значимая робастная модель, обладающая экономическим смыслом.

Далее рассчитаем прогнозные значения увеличивая на 10% каждый фактор (таблица 28):

Таблица 28– Прогноз для Омской области

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Регион** | **x4\*** | **x10** | **y mod** | **y mod\*** | **D** | **y** |
| Омская область | 30008,05 | 2132,912 | 37771,87 | 39143,5 | 1371,632 | 42965,6 |
| **Регион** | **x4** | **x10\*** | **y mod** | **y mod\*** | **D** | **y** |
| Омская область | 27280,04 | 2346,203 | 37771,87 | 41076,46 | 3304,585 | 42965,6 |
| **Регион** | **x4\*** | **x10\*** | **y mod** | **y mod\*** | **D** | **y** |
| Омская область | 30008,05 | 2346,203 | 37771,87 | 42448,09 | 4676,217 | 42965,6 |

При увеличении оборота средних организаций Омской области на 10%, инвестиции в основной фонд выросли на 1371,6 рублей на душу населения, а при увеличении ликвидации основных фондов на 10% - инвестиции выросли на 3304,6 рублей на душу населения, если увеличить оба показателя, то инвестиции увеличатся на 4676,2 рубля на душу населения. Модельное значение данного региона меньше истинного на 5193 рубля, можно предположить, что при увеличении данных показателей на соответствующие значения инвестиции в основной фонд составят 47641,8 рублей на душу населения.

Минимальное изменение при увеличении оборота средних организаций на 10% составило 107,8 рублей, а при увеличении ликвидации основных фондов – 1658,8 рублей, максимальное изменение составило 5359,3 рублей и 32826,8 рублей соответственно. В среднем при увеличении оборота средних организаций на 10%, инвестиции выросли на 1255,2 рублей, а при увеличении ликвидации основных фондов – на 5840,1 рублей на душу населения.

**Заключение**

В данной курсовой работе была построена модель, удовлетворяющая всем требованиям метода главных компонент. Среди всех признаков были отобраны всего два – оборот средних организаций на душу населения и ликвидация основных фондов на душу населения. Полученная эконометрическая модель не нарушает экономический смысл – при росте оборота средних организаций, инвестиции в основной фонд также растут, аналогично с ликвидациями основных фондов.

Два признака были исключены по причине мультиколлинеарности – ВРП и среднедушевые денежные доходы, 6 признаков были исключены по причине незначимости коэффициентов регрессии – уровень занятости населения, численность ИП, преступность, индекс производительности, число реализованных турпакетов и объем розничной торговли.

Вследствие робастного статистического оценивания было исключено три региона – г. Москва, Краснодарский край и Амурская область. После обнаружения гетероскедастичности, было принято решение исключить также еще три региона – Сахалинскую область, Чукотский и Ханты-Мансийский АО.

Метод главных компонент показывал более высокие коэффициенты детерминации, данный метод – хороший инструмент для нивелирования проблемы мультиколлинеарности, однако после преобразования модели, построено на главных компонентах, к исходному признаковому пространству было выяснено, что все модели нарушали экономический смысл, после чего было принято решение отказаться от данного метода.

Также с помощью метода МГК было осуществлено робастное оценивание, выбросы, определенные во главе робастного оценивания были визуально подтверждены на диаграммах рассеивания по первым трем главным компонентам. Также было замечено, что численность ИП, общее количество преступлений и количество реализованных турпакетов довольно слабо влияют на инвестиции в основной фонд.

За счет того, что исходные панельные данные были свернуты по временному признаку, проблемы автокорреляции заведомо не могло быть, что и подтвердил тест Дарбина-Уотсона, также данный тест подтвердил гомоскедастичность остатков выбранной модели.

Стандартная ошибка модели составила 17973 рублей на душу населения. При увеличении каждого признака на 10% инвестиции в основной фонд в среднем увеличились на 7095,3 рублей на душу населения. Дисперсия отобранных признаков объясняет дисперсию инвестиций в основной фонд на 86,4%.

**Источники**

1. Н.П.Тихомиров, Т.М.Тихомирова, О.С.Ушмаев - Методы эконометрики и многомерного статистического анализа – ЗАО «Издательство «Экономика», 2011
2. О.А. Кузнецова, М.С. Татарникова – Эконометрическое моделирование – Самара, издательство СГАУ, 2012
3. Эконометрическое моделирование и прогнозирование социально-экономических процессов – Грибакина Н.М., Бураев Е.В. – «Экономика и социум», 2015
4. «Как работает метод главных компонент (PCA) на простом примере» [Электронный ресурс]. - Режим доступа: <https://habr.com/ru/articles/304214/>
5. «Метод главных компонент (Principal component analysis)» [Электронный ресурс]. - Режим доступа: https://wiki.loginom.ru/articles/principal-component-analysis.html
6. Газета.ру - Когда выходят на пенсию в разных странах мира [Электронный ресурс]. - Режим доступа: <https://www.gazeta.ru/infographics/kogda_vyhodyat_na_pensiyu_v_raznyh_stranah_mira.shtml>
7. Пример построения мультипликативной модели [Электронный ресурс]. - Режим доступа: <https://axd.semestr.ru/dinam/example-multiplicative.php>
8. Эконометрика – В.В.Домбровский – ТГУ, Томск, 2016г.
9. ФСГС. Статистика [Электронный ресурс]. - Режим доступа: https://rosstat.gov.ru/statistic
10. Builtin - A Step-by-Step Explanation of Principal Component Analysis (PCA) [Электронный ресурс]. – Режим доступа: <https://builtin.com/data-science/step-step-explanation-principal-component-analysis>
11. ФСГС. Инвестиции в основной капитал [Электронный ресурс]. - Режим доступа: <https://26.rosstat.gov.ru/folder/42037>
12. ФСГС. Трудовые ресурсы, занятость и безработица [Электронный ресурс]. - Режим доступа: https://rosstat.gov.ru/labour\_force
13. ФСГС. Национальные счета [Электронный ресурс]. - Режим доступа: https://rosstat.gov.ru/statistics/accounts
14. ФСГС. МАЛОЕ И СРЕДНЕЕ ПРЕДПРИНИМАТЕЛЬСТВО В РОССИИ 2022 [Электронный ресурс]. - Режим доступа: <https://rosstat.gov.ru/storage/mediabank/mal_pred_2022.pdf>
15. ФСГС. Уровень жизни [Электронный ресурс]. - Режим доступа: <https://rosstat.gov.ru/folder/13397>
16. Портал правовой статистики. Показатели преступности России [Электронный ресурс]. - Режим доступа: <http://crimestat.ru/offenses_map>
17. ФСГС. Динамика производительности труда [Электронный ресурс]. - Режим доступа: https://rosstat.gov.ru/bgd/regl/b13\_01/IssWWW.exe/Stg/d09/2-1-3.htm
18. ФСГС. Туризм [Электронный ресурс]. - Режим доступа: <https://rosstat.gov.ru/statistics/turizm>
19. T-критерий для независимых выборок - [Электронный ресурс]. - Режим доступа: <https://www.ibm.com/docs/ru/spss-statistics/saas?topic=tests-independent-samples-t-test>
20. ВШЭ. Лекция 7.12.20, часть 2. Гетероскедастичность – Демидова О.А. - [Электронный ресурс]. - Режим доступа https://www.hse.ru/mirror/pubs/share/423684138.pdf