

ФЕДЕРАЛЬНОЕ АГЕНТСТВО СВЯЗИ

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ТЕЛЕКОММУНИКАЦИЙ ИМ. ПРОФ. М.А. БОНЧ-БРУЕВИЧА»
(СПбГУТ)**

**Факультет Информационных систем и технологий
Кафедра Автоматизации предприятий связи**

Системный анализ и принятие решений

Отчет по лабораторной работе №1

**«Прогнозирование объема продаж продукции с помощью моделей и
методов регрессионного анализа»**

Выполнил:
Студент гр. ИСТ-741
Иванов И.И.

Проверил:
ассистент Банцер Е.А.

Санкт-Петербург
2019

Цель выполнения работы

Цель выполнения лабораторной работы – освоение методологии краткосрочного прогнозирования объема продаж продукции с использованием моделей и методов регрессионного анализа и программы Statistica.

Постановка задачи

Имеется совокупность результатов наблюдений за поведением переменной Y в зависимости от изменения одной или нескольких независимых переменных X (X_1, X_2, \dots, X_n). Необходимо установить количественную взаимосвязь между показателем Y и факторами X , т.е. определить такую функциональную зависимость $Y^* = f(X_1, X_2, \dots, X_n)$, которая наилучшим образом описывает имеющиеся экспериментальные данные. На основании построенного уравнения регрессии требуется спрогнозировать значение зависимой переменной Y на шаг вперед (момент времени $(t+1)$) при условии, что значения влияющих факторов на этот период известны.

Описание метода решения задачи

Математическое уравнение, которое описывает линию простой (парной) линейной регрессии с учетом влияния одного фактора, имеет вид:

$$Y_{t+1} = b_0 + b_1 X,$$

где Y_{t+1} – прогнозное значение зависимой переменной на момент времени $(t+1)$;

b_0, b_1 – параметры, которые оцениваются на основе статистических данных (угловые коэффициенты или коэффициенты регрессии);

X – значение влияющего фактора (независимая переменная).

Однофакторная линейная регрессионная модель может быть расширена путем включения в нее более одной независимой переменной. При совместном влиянии на Y нескольких факторов (X_1, X_2, \dots, X_n), уравнением множественной регрессии принимает вид:

$$Y_{t+1} = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_n X_n,$$

где n – число факторов.

Коэффициенты регрессии представляют собой независимые вклады каждой независимой переменной в предсказание зависимой переменной. Если коэффициент b положителен, то связь переменной с зависимой переменной положительна, если коэффициент отрицателен, то и связь носит отрицательный характер (чем меньше значение фактора, тем больше значение переменной Y). Если $b=0$, то связь между переменными отсутствует. Для проверки гипотезы о нулевых значениях коэффициентов регрессии (т.е. об отсутствии связи между Y и совокупностью факторов) анализируются значения F -статистики Фишера. F -критерий определяется отношением дисперсии оценки модели к дисперсии остатка и равен:

$$F = \frac{SSR/q}{SSE/(n-(q+1))},$$

где SSR – сумма квадратов, объясненная уравнением регрессии (Sum of Squares about Regression);

SSE – сумма квадратов остатков (Sum of Squares Errors);

n – число наблюдений;

q – число коэффициентов регрессии.

Гипотеза об отсутствии линейной зависимости между переменной Y и факторами Хотклоняется при больших значениях F -критерия и значении p -level меньше 0,05 (вероятность ошибочной оценки относительно принятой гипотезы не превышает 5% уровня).

Наиболее простым методом определения коэффициентов регрессии является метод наименьших квадратов (МНК). С помощью этого метода параметры регрессионной модели вычисляются таким образом, чтобы сумма квадратов ошибок (расстояний от линии регрессии до фактических значений данных) была бы минимальной.

Функция ошибки при этом равна:

$$f = (b_0 + b_1x_{11} + b_2x_{21} + \dots + b_kx_{k1} - y_1)^2 + (b_0 + b_1x_{12} + b_2x_{22} + \dots + b_kx_{k2} - y_2)^2 + \dots + (b_0 + b_1x_{1n} + b_2x_{2n} + \dots + b_kx_{kn} - y_n)^2$$

Минимизируя функцию f положим:

$$\frac{\partial f}{\partial b_0} = \frac{\partial f}{\partial b_1} = \dots = \frac{\partial f}{\partial b_n} = 0.$$

Для определения коэффициентов модели множественной линейной регрессии, используя систему уравнений, получим систему нормальных линейных уравнений, которая в векторно-матричной форме имеет вид:

$$\begin{bmatrix} n & \sum_{i=1}^n x_{1i} & \sum_{i=1}^n x_{2i} & \dots & \sum_{i=1}^n x_{ki} \\ \sum_{i=1}^n x_{1i} & \sum_{i=1}^n x_{1i}^2 & \sum_{i=1}^n x_{1i}x_{2i} & \dots & \sum_{i=1}^n x_{1i}x_{ki} \\ \sum_{i=1}^n x_{2i} & \sum_{i=1}^n x_{2i}x_{1i} & \sum_{i=1}^n x_{2i}^2 & \dots & \sum_{i=1}^n x_{2i}x_{ki} \\ \dots & \dots & \dots & \dots & \dots \\ \sum_{i=1}^n x_{ki} & \sum_{i=1}^n x_{ki}x_{1i} & \sum_{i=1}^n x_{ki}x_{2i} & \dots & \sum_{i=1}^n x_{ki}^2 \end{bmatrix} \cdot \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \dots \\ b_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{1i}y_i \\ \sum_{i=1}^n x_{2i}y_i \\ \dots \\ \sum_{i=1}^n x_{ki}y_i \end{bmatrix},$$

где n - число экспериментальных точек;

i - номер точки.

Отклонение отдельной точки от линии регрессии (предсказанного значения) называется остатком. Чем меньше разброс значений остатков около линии регрессии по отношению к общему разбросу значений, тем лучше прогноз. Оценка качества линейной регрессии проводится с помощью коэффициента детерминации R^2 , который показывает какая доля дисперсии отклика объясняется влиянием независимых переменных в построенной модели.

$$R^2 = SSR / SST, \text{ где}$$

SST – полная сумма квадратов (Total Sum of Squares).

Если связь между переменными X и Y отсутствует, то отношение остаточной изменчивости переменной Y к исходной дисперсии равно 1. Если X и Y коррелируют между собой, то остаточная изменчивость отсутствует и отношение дисперсий будет равно 0. Например, если имеется $R^2 = 0,4$, то изменчивость значений переменной Y около линии регрессии составляет 1-0,4 от исходной дисперсии, т.е. 40% от исходной изменчивости могут быть объяснены, а 60% остаточной изменчивости остаются необъясненными. Значение R^2 является индикатором степени подгонки модели к данным (значение R^2 близкое к

Ипоказывает, что модель объясняет почти всю изменчивость соответствующих переменных). При поиске лучшей регрессионной модели руководствуются требованием $R^2 \geq 0,8$.

Функциональные возможности программы Statistica

Для решения задачи краткосрочного прогнозирования объема продаж предполагается использовать программу Statistica. Пакет прикладных программ, разработанный компанией StatSoft, позволяет проводить исчерпывающий, всесторонний анализ данных, представлять результаты анализа в виде таблиц и графиков, автоматически создавать отчеты о проделанной работе. Предоставляет мощные и удобные в использовании инструменты для статистического и графического анализа, реализует функции управления данными, добычи и визуализации данных, datamining и др.

Программа Statistica имеет модульную структуру, т.е. состоит из модулей, каждый из которых используется для решения конкретного класса задач, а именно: анализ временных рядов и прогнозирование, множественная регрессия, нелинейное оценивание, факторный анализ, кластерный анализ, канонический анализ, непараметрическая статистика, дисперсионный и дискриминантный анализ. Несколько модулей объединены в группу промышленная статистика: контроль качества, анализ процессов, планирование эксперимента.

Оценка коэффициентов однофакторной и многофакторной линейной регрессии осуществляется в отдельном окне системы Statistica, где представлены коэффициенты, оцененные методом наименьших квадратов, коэффициент детерминации, статистика Фишера оценки значимости регрессии, статистики Стьюдента, оценки значимости коэффициентов, коэффициент корреляции (матрица корреляций), статистика Дарбина-Уотсона. Можно анализировать большие модели, содержащие до 500 переменных.

Пример (вариант 1)

Описание деловой ситуации

Пусть предприятие работает на рынке определенного продукта. При формировании маркетингового решения возникла необходимость в прогнозировании уровня платежеспособного спроса (объема продаж) на выпускаемую продукцию.

Для решения задачи краткосрочного прогнозирования спроса предполагается использовать модели и методы регрессионного анализа.

Маркетологи фирмы располагают статистическими данными за период, равный 30 месяцам, о фактических значениях объемов продаж продукции по месяцам, расходах на рекламу по месяцам, ценах на продукцию фирмы и на продукцию конкурирующей фирмы (см. табл. 1). Такие факторы, как расходы на рекламу, цена на продукцию фирмы и ее основного конкурента были выбраны как наиболее значимые по степени влияния на выходной показатель – объем продаж продукции.

Требуется: 1. Рассчитать ожидаемый предприятием объем продаж продукции на 31-й месяц работы при условии, что предполагаемые на этот месяц значения факторов, влияющих на объем продаж, составят: расходы на рекламу – 93 тыс. руб., цена единицы продукции – 340 руб., цена единицы продукции конкурирующей фирмы – 343 руб.

2. Построить последовательно однофакторную линейную регрессионную модель (с учетом только одного фактора – расходов на рекламу), двухфакторную линейную регрессионную модель (с учетом таких факторов, как расходы на рекламу и цена единицы продукции предприятия) и трехфакторную линейную регрессионную модель (с учетом трех

факторов - расходы на рекламу, цена единицы продукции предприятия и цена на единицу продукции фирмы-конкурента).

3. Рассчитать оценки прогнозов объема продаж в 31 месяце, используя построенные регрессионные модели.

4. Провести сравнительный анализ прогнозных оценок объемов продаж продукции, полученных с помощью однофакторной, двухфакторной и трехфакторной регрессионных моделей.

При поиске лучшей регрессионной модели следует руководствоваться следующими наиболее общими требованиями:

1. Регрессионная модель должна объяснять не менее 80% вариации зависимой переменной, т.е. $R^2 \geq 0.8$.
2. Стандартная ошибка оценки зависимой переменной по уравнению должна составлять не более 5% среднего значения зависимой переменной;
3. Коэффициенты уравнения регрессии и его свободный член должны быть значимы на 5%-ом уровне.
4. Остатки от регрессии должны быть нормально распределены и без систематической составляющей.

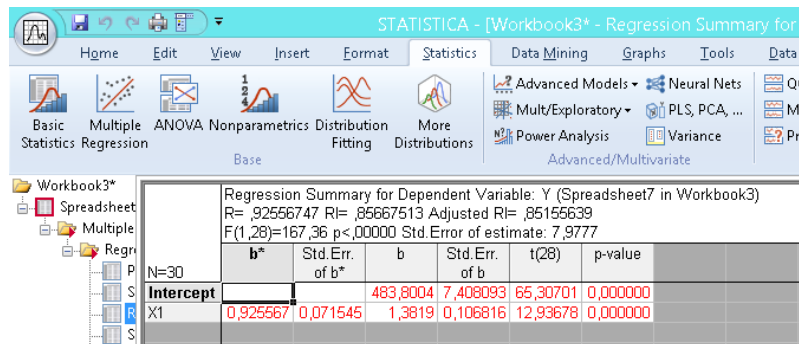
Таблица 1. Исходные данные для прогнозирования спроса

| Номер месяца (t) | Фактический объем продаж за месяц (тыс.руб.) (Y) | Расходы на рекламу за месяц (тыс. руб.) (X ₁) | Цена продукта (руб./ед. продукта) (X ₂) | Цена продукта фирмы-конкурента (руб./ед. продукта) (X ₃) |
|---------------------|---|--|---|---|
| 1 | 545 | 58 | 287 | 290 |
| 2 | 549 | 52 | 289 | 291 |
| 3 | 545 | 50 | 290 | 292 |
| 4 | 550 | 51 | 279 | 295 |
| 5 | 562 | 53 | 278 | 293 |
| 6 | 568 | 49 | 283 | 296 |
| 7 | 565 | 53 | 291 | 295 |
| 8 | 568 | 59 | 293 | 295 |
| 9 | 564 | 60 | 293 | 297 |
| 10 | 553 | 61 | 291 | 293 |
| 11 | 562 | 57 | 289 | 292 |
| 12 | 560 | 55 | 294 | 295 |
| 13 | 554 | 62 | 299 | 302 |
| 14 | 581 | 68 | 301 | 304 |
| 15 | 585 | 67 | 301 | 304 |
| 16 | 587 | 75 | 299 | 305 |
| 17 | 580 | 63 | 315 | 318 |
| 18 | 584 | 64 | 318 | 322 |
| 19 | 586 | 69 | 313 | 320 |
| 20 | 585 | 70 | 302 | 308 |
| 21 | 583 | 75 | 321 | 327 |
| 22 | 589 | 74 | 334 | 339 |
| 23 | 591 | 79 | 328 | 330 |
| 24 | 595 | 80 | 320 | 325 |
| 25 | 600 | 83 | 329 | 334 |
| 26 | 605 | 85 | 330 | 337 |
| 27 | 608 | 89 | 337 | 339 |
| 28 | 610 | 92 | 338 | 340 |
| 29 | 612 | 94 | 340 | 346 |
| 30 | 607 | 93 | 339 | 345 |
| 31 | ? | 93 | 340 | 343 |

Анализ результатов решения задачи

В программе Statistica задаются исходные данные по 30 месяцам из таблицы 1.

1. Проводится построение однофакторной линейной регрессионной модели (с учетом одного фактора – расходов на рекламу) вида $Y=b_0+b_1X_1$. Коэффициенты регрессии b_0 , b_1 , рассчитанные для однофакторной модели представлены на рисунке 1.



| | b* | Std. Err. of b* | b | Std. Err. of b | t(28) | p-value |
|-----------|----------|-----------------|----------|----------------|----------|----------|
| Intercept | | | 483,8004 | 7,408093 | 65,30701 | 0,000000 |
| X1 | 0,925567 | 0,071545 | 1,3819 | 0,106816 | 12,93678 | 0,000000 |

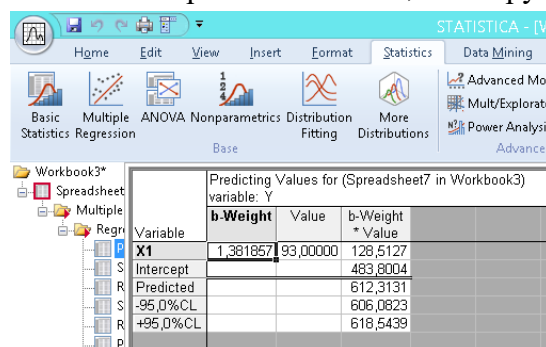
Рисунок 1. Итоги регрессионного анализа однофакторной модели

Проведен расчет значений коэффициентов регрессии для модели с учетом влияния фактора «Расходы на рекламу». Выражение однофакторной линейной регрессии можно представить в виде: $Y_{31}=483,8+1,38*X_1$.

Расчет прогнозного значения объема продаж на 31-й месяц в ручном режиме:

$$Y_{31}=483,8+1,38*93=612,14$$

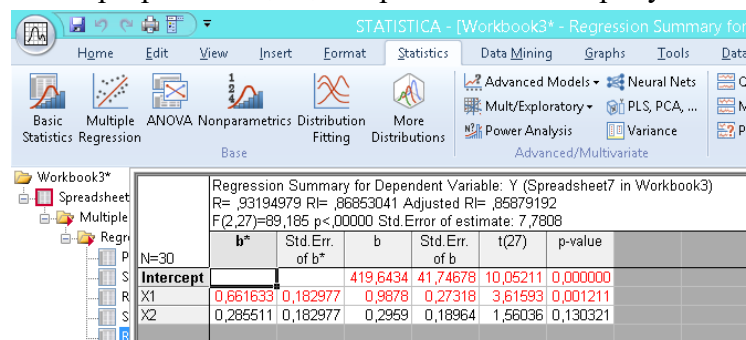
Прогнозное значение объема продаж на 31-й месяц с учетом влияния расходов на рекламу, полученное в автоматическом режиме $Y_{31}=612,31$ тыс.руб. (рис. 2).



| Variable | b-Weight | Value | b-Weight * Value |
|-----------|----------|----------|------------------|
| X1 | 1,381857 | 93,00000 | 128,5127 |
| Intercept | | | 483,8004 |
| Predicted | | | 612,3131 |
| -95,0%CL | | | 606,0823 |
| +95,0%CL | | | 618,5439 |

Рисунок 2. Расчет прогнозного значения объема продаж с использованием однофакторной линейной регрессионной модели

2. Строится двухфакторная линейная регрессионная модель (с учетом двух факторов – расходов на рекламу и цены единицы продукции) вида $Y=b_0+b_1X_1+b_2X_2$. Значения коэффициентов регрессии b_0 , b_1 , b_2 представлены на рисунке 3.



| | b* | Std. Err. of b* | b | Std. Err. of b | t(27) | p-value |
|-----------|----------|-----------------|----------|----------------|----------|----------|
| Intercept | | | 419,6434 | 41,74678 | 10,05211 | 0,000000 |
| X1 | 0,661633 | 0,182977 | 0,9878 | 0,27318 | 3,61593 | 0,001211 |
| X2 | 0,285511 | 0,182977 | 0,2959 | 0,18964 | 1,56036 | 0,130321 |

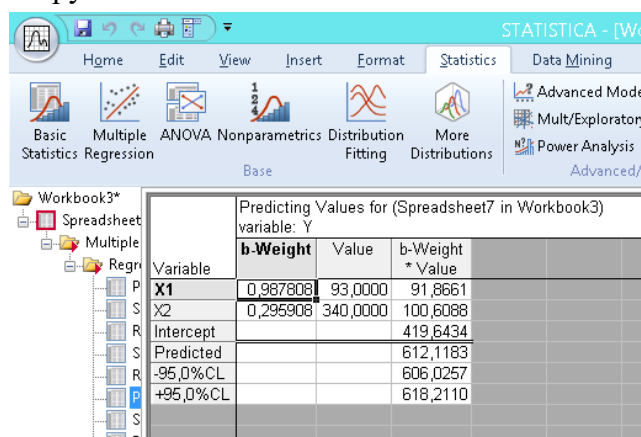
Рисунок 3. Итоги регрессионного анализа двухфакторной модели

Проведен расчет значений коэффициентов регрессии для модели с учетом влияния факторов «Расходы на рекламу» и «Цена единицы продукции». Выражение двухфакторной линейной регрессии можно представить в виде: $Y_{31}=419,64+0,99*X_1+0,3*X_2$.

Расчет прогнозного значения объема продаж на 31-й месяц в ручном режиме:

$$Y_{31}=419,64+0,99*93+0,3*340=613,71$$

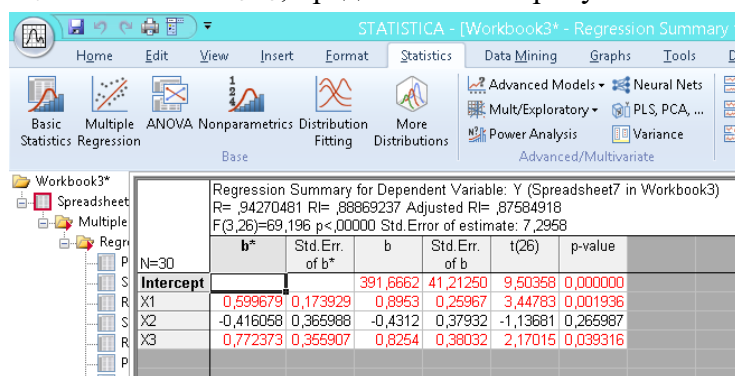
Прогнозное значение объема продаж на 31-й месяц с учетом влияния расходов на рекламу и цены единицы продукции предприятия, полученное в автоматическом режиме составляет $Y_{31}=612,12$ тыс.руб.



| Variable | b-Weight | Value | b-Weight * Value |
|-----------|----------|----------|------------------|
| X1 | 0,987808 | 93,0000 | 91,8661 |
| X2 | 0,295908 | 340,0000 | 100,6088 |
| Intercept | | | 419,6434 |
| Predicted | | | 612,1183 |
| -95,0%CL | | | 606,0257 |
| +95,0%CL | | | 618,2110 |

Рисунок 4. Расчет прогнозного значения объема продаж при использовании двухфакторной регрессионной модели

3. Значения коэффициентов регрессии b_0 , b_1 , b_2 , b_3 , рассчитанные для трехфакторной линейной регрессионной модели (с учетом трех факторов – расходов на рекламу, цены единицы продукции предприятия и цены единицы продукции фирмы-конкурента) вида $Y=b_0+b_1X_1+b_2X_2+b_3X_3$, представлены на рисунке 5.



| | N | b* | Std. Err. of b* | b | Std. Err. of b | t(26) | p-value |
|-----------|----|-----------|-----------------|---------|----------------|----------|----------|
| Intercept | 30 | 391,6662 | 41,21250 | 9,50358 | 0,000000 | | |
| X1 | | 0,599679 | 0,173929 | 0,8953 | 0,25967 | 3,44783 | 0,001936 |
| X2 | | -0,418058 | 0,365988 | -0,4312 | 0,37932 | -1,13681 | 0,265967 |
| X3 | | 0,772373 | 0,355907 | 0,8254 | 0,38032 | 2,17015 | 0,039316 |

Рисунок 5. Итоги регрессионного анализа трехфакторной модели

Проведен расчет значений коэффициентов регрессии для модели с учетом влияния факторов «Расходы на рекламу», «Цена единицы продукции предприятия» и «Цена единицы продукции фирмы-конкурента». Выражение трехфакторной линейной регрессии можно представить в виде: $Y_{31}=391,67+0,895*X_1-0,43*X_2+0,83*X_3$

Расчет прогнозного значения объема продаж на 31-й месяц в ручном режиме:

$$Y_{31}=391,67+0,895*93-0,43*340+0,83*343=613,395$$

| Variable | b-Weight | Value | b-Weight * Value |
|-----------|-----------|----------|------------------|
| X1 | 0,895311 | 93,0000 | 83,264 |
| X2 | -0,431210 | 340,0000 | -146,611 |
| X3 | 0,825356 | 343,0000 | 283,097 |
| Intercept | | | 391,666 |
| Predicted | | | 611,416 |
| -95,0%CL | | | 605,654 |
| +95,0%CL | | | 617,178 |

Рисунок 6. Расчет прогнозного значения объема продаж с использованием трехфакторной регрессионной модели

Прогнозное значение объема продаж на 31-й месяц с учетом влияния расходов на рекламу, цены единицы продукции предприятия и цены единицы продукции фирмы-конкурента, полученное в автоматическом режиме $Y_{31}=611,42$ тыс.руб.

Заключение

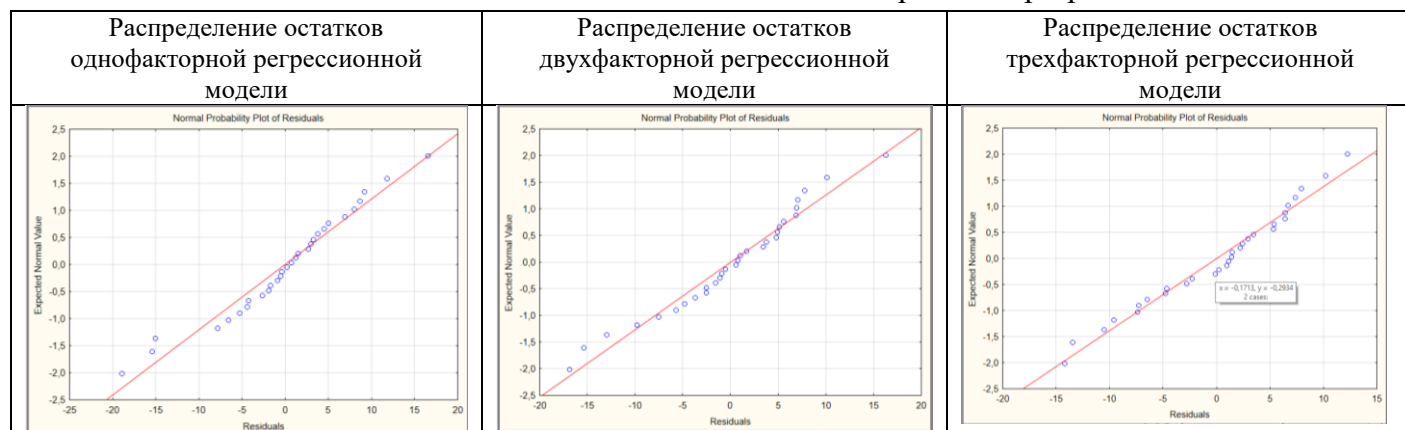
В ходе лабораторной работы была решена задача прогнозирования объема продаж с применением метода регрессионного анализа. Осуществлено последовательное построение моделей регрессионного анализа с учетом влияния одного, двух и трех рассматриваемых факторов на значение объема продаж для 31 месяца работы предприятия. Анализ построенных моделей показал, что все три модели корректны, имеют значения $R^2 > 0,85$, т.е. объясняют больше 85% разброса значений переменной Y относительно среднего:

- для однофакторной модели $R^2=0,857$;
- для двухфакторной модели $R^2=0,869$;
- для трехфакторной модели $R^2=0,889$.

Во всех построенных регрессионных моделях стандартная ошибка оценки зависимой переменной составляет порядка 2,5%, что является допустимой нормой. Построенные регрессии значимы, а гипотеза об отсутствии связи между переменными может быть отклонена, т.к. большим значениям F -критериев соответствуют уровни значимости (p -level) меньше 5%. Результаты получены на последнем шаге регрессии и проведен анализ остатков для каждой регрессионной модели (табл. 2). По графикам можно сделать вывод, что остатки нормально распределены (в пределах ± 18 ед.), заметных выбросов нет.

Фактор X_2 (цена единицы продукции предприятия) в двухфакторной и трехфакторной моделях имеет низкий уровень значимости (p -level больше 0,05), т.е. этот фактор в меньшей степени влияет на изменение уровня объема продаж, чем остальные факторы.

Таблица 2. Распределение остатков, полученных по результатам построенных регрессионных моделей



Таким образом, каждая из построенных регрессионных моделей может быть использована для решения задачи прогнозирования объема продаж продукции предприятия. Окончательным решением задачи прогнозирования будем считать прогнозное значение $Y_{31}=611,42$ тыс.руб., полученное при использовании трехфакторной регрессионной модели, т.к. этой модели соответствует наиболее высокий уровень коэффициента детерминации R^2 и наименьший разброс остатков (в пределах ± 15 ед.).