УДК 519.233.5

ПРИМЕНЕНИЕ МАТЕМАТИЧЕСКИХ МЕТОДОВ В ИССЛЕДОВАНИИ И АНАЛИЗЕ СТАТИСТИЧЕСКИХ ДАННЫХ, НА ПРИМЕРЕ ПОСТРОЕНИЯ МОДЕЛЕЙ КУРСОВ ВАЛЮТ

Рыспаев А.О., Абдиева Л.К., Сыдыкова А.Ж., Жапаркулов Ж.Ш.
КГУСТА им. Н. Исанова

В работе построены и исследованы математические модели статистических экономических данных - курсов валют с применением корреляционного и регрессионного анализов.

***Ключевые слова**:* математические модели, корреляционный анализ, регрессионный анализ.

СТАТИСТИКАЛЫК МААЛЫМАТТАРДЫ ИЗИЛДӨӨ ЖАНА ТАЛДООДО МАТЕМАТИКАЛЫК МЕТОДДАРДЫ КОЛДОНУУ, ВАЛЮТА КУРСТАРЫНЫН МОДЕЛДЕРИН ТҮЗҮҮНҮН МИСАЛЫНДА

Рыспаев А.О., Абдиева Л.К., Сыдыкова А.Ж., Жапаркулов Ж.Ш.
Н.Исанов ат. КМКТАУ

Макалада статистикалык экономикалык маалыматтардын - валюта курсунун корреляциялык жана регрессиялык анализдерди колдонуу менен математикалык моделдери түзүлгөн жана изилденген.

***Баштапкы сөздөр*:** математикалык моделдер, корреляциялык анализ, регрессиялык анализ.

APPLICATION OF MATHEMATICAL METHODS IN THE STUDY AND ANALYSIS OF STATISTICAL DATA, USING THE EXAMPLE OF BUILDING MODELS OF EXCHANGE RATES

Ryspaev A.O., Abdieva L.K., Sydykova A.J., Japarkulov J.Sh.
KSUCTA named of N. Isanova

Mathematical models of statistical economic data - exchange rates with the use of correlation and regression analyses are constructed and investigated in the work.
Keywords: mathematical models, correlation analysis, regression analysis.

В процессе своей жизнедеятельности человечество еще с древних времен осуществляет учет так называемых статистических данных о самых разных явлениях, предметах и т.д. в самых разнообразных областях (экономике, биологии, медицине и др.). В настоящее время под термином «статистические данные» понимают все собранные сведения, которые в дальнейшем подвергаются статистической обработке и анализу. Анализ данных нельзя рассматривать только как обработку информации после ее сбора [1]. Анализ данных — это прежде всего средство проверки гипотез и решения задач исследователя, в качестве гипотезы в анализе данных часто выступает предположение о влиянии какого-либо фактора или группы факторов на результат.

Статистические методы анализа и обработки данных по своей сути основаны на достижениях математики, в частности математической статистики [2]. Выбор конкретных методов зависит от поставленных исследовательских задач, особенностей и специфики изучаемых процессов. Наиболее часто используемыми методами математической статистики в исследованиях являются корреляционный и регрессионный анализ [3].

Корреляционный анализ показывает тип связи (положительная или отрицательная, прямая или обратная) и измеряет тесноту связи между наблюдениями, которые являются случайными и выбранными из некоторой генеральной совокупности данных. Регрессионный анализ метод установления аналитического выражения стохастической зависимости между исследуемыми признаками. Уравнение регрессии показывает, как в среднем меняется результативный показатель у при изменении любого из независимых показателей х. В отличие от

корреляционного анализа, который только отвечает на вопрос, существует ли связь между признаками, регрессионный анализ дает ее формализованное выражение.

Разработка математической модели включает несколько этапов:

1) Формулировка цели исследования. На этом этапе нужно определить, к примеру экономические объекты, явления или процессы, между которыми необходимо найти зависимость. Нужно установить период исследования. Нужно определиться с классификацией результирующего фактора (отклика) и независимых объясняющих факторов (регрессоров). Т.е. необходимо сформулировать гипотезы о зависимости экономических явлений, причем они должны иметь экономически приемлемый смысл.

2) Сбор и обработка статистических данных. Чаще всего экономические данные структурированы по времени и даются в виде табличной формы. Нужно определиться с объемом выборки. Рекомендуется чтобы получить статистически значимые модели более 30 наблюдений для одного фактора. В случае рассмотрения нескольких факторов, количество данных увеличивают в 3-5 раз.

3) Выделение факторов, наиболее значимо оказывающих воздействие на результирующую функцию. Для этого определим степени взаимосвязей между факторами по значениям коэффициентов корреляции:

$$r_{X_u X_v} = \frac{\overline{X_u X_v} - \overline{X_u}\,\overline{X_v}}{\sigma(X_u)\sigma(X_v)},$$

v, u – порядковые номера факторов, n – объем выборки.

Исследования указывают, если коэффициент корреляции меньше 0.7 то соответствующий фактор можно исключить из модели [4].

4) Выбор функции, построение уравнения. На этом этапе формулируются гипотезы о виде зависимости результирующего фактора от независимых факторов (линейная или нелинейная, однофакторная или множественная и т.д.). При выборе функции лучшим критерием считают – простоту функции, т.е. чем проще функция и чем меньше в ней задействовано

факторов, тем она лучше отражает регрессионную зависимость. В данной работе будут рассмотрены линейные однофакторные модели.

5) Расчет параметров уравнения. Расчет параметров уравнения регрессии будет осуществлен на основе метода наименьших квадратов, являющимся одним из базовых методов регрессионного анализа. Он основан на минимизации суммы квадратов отклонений фактических значений отклика от расчетных - определяемых по уравнению регрессии, т.е. это минимизация квадратов остатков регрессии:

$$\sum_i e_i^2 = \sum_i (y_i - f_i(x))^2 \to \min$$

.

Фактически решается задача оптимизации, для этого находят стационарные точки функции, приравнивая частные производные функции по ее параметрам к нулю и решают полученную систему уравнений.

6) оценка статистической значимости, качества и адекватности полученной модели.

Коэффициенты регрессии проходят проверку на статистическую значимость, обычно по критерию Стьюдента: $t_k = \frac{a_k}{\sigma_{a_k}}$,

a_k - коэффициент регрессии при к-том факторе, σ_{a_k} - стандартное отклонение оценки параметра a_k. Если при выбранном уровне значимости α вычисленная статистика больше критического значения, говорят, что коэффициент является статистически значимым. В противном случае, коэффициент является статистически незначимым и его значение можно взять равным нулю для данной модели. На практике рекомендуется применять модели со статистически значимыми коэффициентами.

Для оценки качества модели применяют коэффициент детерминации R^2. Значение коэффициента находится между 0 и 1. Если R^2 равен нулю, то считается что регрессоры не оказывают влияние на

отклик, не улучшают качество предсказания отклика. Если R^2 равен единице, то все точки наблюдений находятся на линии регрессии, и все изменения (вариации) результирующей функции объяснены изменениями (вариациями) независимых факторов – регрессоров. Таким образом, чем ближе R^2 к единице, тем лучше качество построенной модели. Коэффициент детерминации определяют по формуле:

$$R^2 = 1 - \frac{ESS}{TSS} = \frac{RSS}{TSS},$$

где $TSS = \sum (Y_t - \bar{Y})^2, ESS = \sum (Y_t - \bar{\bar{Y}}_t)^2, RSS = \sum (\bar{\bar{Y}}_t - \bar{Y})^2,$ причем $TSS = RSS + ESS$.

Далее осуществляют проверку адекватности и статистической значимости самой регрессионной модели по критерию Фишера:

$$F = \frac{MSE}{\sigma_e^2},$$

где MSE (Mean Square Error) есть среднеквадратическая ошибка (остаточная дисперсия) модели, часто обозначается еще и как ESS, характеризует среднее отклонение вычисленных и наблюдаемых значений отклика, σ_e^2 - среднеквадратическая ошибка воспроизводимости отклика. Вообще, если MSE > σ_e^2, считается построенная модель адекватно описывает наблюдаемые данные. Если MSE ≤ σ_e^2, то построенная модель имеет элементы – данные, которые в большей степени изменения наблюдаемых данных описывают случайно.

При выявлении статистически незначимых факторов необходимо произвести корректировку модели удалением не значимых факторов или добавлением новых факторов.

Рассмотрим применение корреляционного и регрессионного анализов для построения математической модели курсов валют доллара и евро через их относительные курсы к сому с марта 2020 года по март 2021 года. Данные взяты с официального сайта НБ КР. Из 365 данных после обработки (исключения дублирующих значений) было оставлено

248 чисел. Нужно построить и исследовать модели обменных курсов валют доллара и евро.

Сформулируем для проверки следующие гипотезы:

1) Первая гипотеза: курс доллара USD (отклик У) следует за курсом евро EUR (регрессор Х);

2) Вторая гипотеза: курс евро EUR (отклик У) следует за курсом доллара USD (регрессор Х).

Введем для удобства записи следующие обозначения:

Usd_i, Usd_{i-1} - курсы котировок доллара в моменты времени i и i-1,

Eur_i, Eur_{i-1} - курсы котировок евро в моменты времени i и i-1,

$$U_i = \ln \frac{Usd_i}{Usd_{i-1}}, \quad E_i = \ln \frac{Eur_i}{Eur_{i-1}}, \quad i = \overline{1,n}, \text{ n=248.}$$

В общем виде линейная однофакторная регрессионная модель записывается в виде:

$$Y = \alpha_0 \cdot X + \alpha_1, \tag{1}$$

где Х – регрессор (независимый фактор), Y – отклик (зависимый фактор) от регрессора Х, α_0, α_1 - неизвестные коэффициенты уравнения регрессии, которые необходимо найти.

Были построены две линейные однофакторные модели на «первичных» данных (1-я и 2-я модели) и две линейные однофакторные модели на прологарифмированных данных (3-я и 4-я модели). Под «первичными» имеются в виду 248 данные курсов валют. Прологарифмированные данные (247 данных) – это «первичные» данные, переработанные по формуле, так называемой логарифмической доходности: $A_i = \ln \frac{A_i}{A_{i-1}}$. Логарифмирование преобразует ассиметричные данные в более симметричные, так как происходит «растягивание» шкалы возле нуля, малые значения, сгруппированные вместе, распределяются вдоль шкалы. Данная процедура минимизирует влияние

асимметричных данных на результаты корреляционного и регрессионного анализа.

Коэффициенты уравнения регрессий и другие требуемые статистики были найдены через пакет прикладных программ «Анализ данных» MS Excel. Результаты следующие:

1) Уравнение регрессии:

1 модель: Usd = 0.49249Eur+34.13255

2 модель: Eur = 1.69023Usd-42.20475

3 модель (логарифмы курсов валют): U = 0.90994E – 0.000285

4 модель (логарифмы курсов валют): E = 0.98274U + 0.000405

2) Коэффициенты корреляции моделей обозначим индексами: r1=0.9124, r2=0.9124, r3=0.9456, r4=0.9456. Коэффициенты корреляции всех моделей получили равными больше 0.9, это говорит о том, что связь между котировками доллара и евро существует сильная. Данные факторы должны присутствовать в регрессии.

3) Оценка статистической значимости, качества и адекватности полученной модели:

а) t – статистики: t11=34.87 (для коэффициента регрессора 1 модели) и t12=26.07 (для коэффициента свободного члена 1 модели), t21=34.88 и t22=10.92, t31=45.41 и t32=1.03426, t41=45.41 и t42=1.41. При уровне значимости α = 0.05 и f = 243 степенями свободы tкрит = 1.96. Сравнивая значения полученных статистик с критическим, получили, что коэффициенты регрессоров во всех четырех моделях статистически значимы. Статистически незначимые оказались коэффициенты свободных членов третьей и четвертой моделей. Так как данные коэффициенты в принципе не несут какую-либо смысловую нагрузку, можно не обращать внимания на их незначимость.

Стандартные ошибки коэффициентов регрессии S11=0.014 S12=0.014, S21=0.048 S22=0.048, S31=0.02 S32=0.02, S41=0.022 S32=0.022 являются небольшими числами. Чем меньше стандартные

ошибки, тем точнее являются оценки коэффициентов регрессии. Поэтому полученные оценки коэффициентов регрессии являются точными и надежными.

б) F – статистика: F1=1217.02, F2=1217.02, F3=2062.81, F4=2062.81, Fтабл = 3.87. Сравнивая значения полученных статистик с табличным, делаем вывод, что построенные модели значимы и адекватны наблюдаемым данным.

в) Коэффициент детерминации: R^2=0,83 (1 и 2 моделей), R^2=0,89 (3 и 4 моделей). Значения R^2 моделей больше 0.8, это означает, что вариации отклика - доллара в среднем описываются вариациями регрессора – евро более чем на 80%, следовательно модели являются качественными.

г) Среднеквадратическая ошибка MSE всех моделей значительно больше среднеквадратической ошибки воспроизводимости отклика σ_e^2 это говорит, что в выборке нет «переполнения» случайными данными и модель описывает наблюдаемые значения удовлетворительно. Остаточные сумма квадратов RSS в третьей и четвертой моделях гораздо меньше, чем у первой и второй.

Математические ожидания остатков Me и дисперсии остатков De для моделей следующие: Me1=0.00062327, De1=0.000003264, Me2=-0.00056398, De2=0.000002647, Me3=0.000000000000000228, De3=0.00000000000000000013, Me4=-0.0000000000000002645, De4=0.00000000000000014. Математические ожидания остатков Me очень близки к нулю, дисперсии остатков De есть конечные числа, значит регрессионные ошибки являются случайными и независимыми друг от друга. Причем, математические ожидания и дисперсии остатков в третьей и четвертой моделях намного меньше соответствующих значений первой и второй моделей. Это свидетельствует о более лучшем отражении и аппроксимации реальных данных регрессионными расчетными по третьей и четвертой моделям.

Таким образом можно отметить, что из представленных результатов, все четыре модели являются адекватными и качественными. Значения остальных статистик, критерия Фишера, коэффициента детерминации, коэффициент корреляции, остаточные суммы квадратов третьей и четвертой моделей лучше, чем у первой и второй моделей. Поэтому можно сделать вывод, что моделями лучше отражающими и аппроксимирующими реальные данные являются третья и четвертая. Выявить из них более предпочтительную не удается, так как обе модели имеют почти одинаковые значения статистик. Поэтому обе сформулированные выше гипотезы имеют место быть в случае однофакторных регрессионных моделей. Вопрос требует дальнейшего исследования и будет продолжен в следующей работе в случае множественных регрессионных моделей. Возможно это даст выявить более предпочтительную модель.

ЛИТЕРАТУРА

1. Рубаков С.В. Современные методы анализа данных //Управление наукой и наукометрия. – 2008. - №7. – с.165-175.
2. Калинин А. Г. Обработка данных методами математической статистики: монография / А. Г. Калинин. – Чита: ЗИП СибУПК, 2015. – 106 с.
3. Тимофеев С.А., Юрьев В.Н. Прогнозирование динамики курсов валют на основе статистического анализа показателей мировой экономики //Научно-техн. ведомости СПбГПУ. -2013, №3(173) - с.16-22.
4. Галустян М.Ж. Проблемы использования метода наименьших квадратов при оценке и прогнозировании динамики фондовых рынков //Известия ТулГУ. – 2015. - №2(1) – с. 88-92.