МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра Математической экономики

Регрессионная модель спроса в интернет-торговле

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студентки 4 курса 451 группы

направление 38.03.05 — Бизнес-информатика

механико-математического факультета

Гудковой Настасьи Владиславовны

Научный руководитель
доцент, к.ф. - м.н. В.В. Новиков

Зав. кафедрой
д.ф.-м.н., профессор С.И. Дудов

Саратов 2019

Введение. Появление нового направления современного бизнеса — электронного бизнеса, связано с развитием сети Интернет. Важнейшим составным элементом данного направления бизнеса является электронная коммерция. Примером может служить электронная торговля, которая представляет собой финансовые операции и сделки, выполняемые через сеть Интернет, в ходе которых совершаются покупки и продажи товаров и услуг, а также переводы денежных средств. Интернет-торговля является крайне перспективным рынком, и для достижения успеха в этой сфере необходимо учитывать особенности, диктуемые поведением интернет-пользователей, а также внешними факторами, которые оказывают непосредственное влияние на развитие данного направления. Анализу развития рынка электронной торговли посвящено множество исследований.

Любой посетитель сети Интернет является потенциальным клиентом множества виртуальных торговых площадок. По подсчетам компаний интернет-торговли (АКИТ) и компании Admitad, структура расходов россиян в интернете в денежном выражении в 2018 году выглядит следующим образом: «электронная техника» (27%), «одежда, обувь и аксессуары» (21%), товары для дома и ремонта (11%), автозапчасти (6%), красота и здоровье (4%), книги (4%), другие категории (4%), товары для детей (3%) и продукты питания (2%).

Численность аудитории виртуальных торговых площадок является одним из показателей уровня развития электронной торговли. В связи с развитием в России информационного общества, внедрением интернет-технологий в различные сферы жизнедеятельности возникает проблема анализа аудитории интернет-магазинов, и возможность прогноза ее расширения приобретают все большее значение.

Целью бакалаврской работы является анализ сезонности российских онлайн-площадок на основе динамики месячной аудитории в 2013-2018 гг. проектов Яндекс.Маркет и Price, используя средства языка R.

Для достижения поставленной цели, необходимо решить несколько задач:

- рассмотреть вопросы регрессионного анализа;
- рассмотреть язык R;

- осуществить анализ сезонности российских онлайн-площадок на основе динамики месячной аудитории в 2013-2017 гг. проекта Яндекс.Маркет и проекта Price.

Основное содержание работы содержит 5 разделов:

1. Введение
2. Математические аспекты эконометрического анализа.
3. Анализ сезонности электронной торговли.
4. Программная реализация метода в среде R.
5. Заключение

Во **введении** формулируется цель работы и решаемые задачи.

В **первом разделе** рассматриваются математические аспекты эконометрического анализа.

Закономерности в экономике выражаются в виде связей и свойств экономических показателей, а также в виде математических моделей их поведения. Такие зависимости и модели могут быть получены только путем обработки реальных статистических данных с учетом внутренних механизмов связи и случайных факторов. Математическая статистика и ее применение в экономики называется эконометрикой.

Эконометрика позволяет строить экономические модели и оценивать их параметры, проверять гипотезы о свойствах экономических показателей и формах их связи.

Статистическую связь между признаками выражают с помощью такой математической функции, которая дает наименьшее отклонение от полученных при наблюдении значений признаков. Уравнение такой функции является уравнением связи между результативными и факториальными признаками.

Регрессионный анализ заключается в определении аналитического выражения связи, в котором изменение одной величины обусловлено влиянием одной или нескольких независимых величин, а множество всех прочих факторов, также оказывающих влияние на зависимую переменную, принимается за постоянные и средние значения.

По форме зависимости различают:

a) линейную регрессию, которая выражается уравнением прямой (линейной) функцией вида: $y = a + bx$;

b) нелинейную регрессию, которая выражается уравнениями степенной, показательной, экспоненциальной функцией, а также уравнениями параболы и гиперболы вида:

- парабола: $y = a + bx + cx^2$
- гипербола: $y = a + bx$ и т.д.

Методом наименьших квадратов (МНК) называется нахождение параметров прямой, удовлетворяющей данному требованию:

$$S = \sum (y_i - \overline{y_i})^2 \to min$$

т.е сумма квадратов отклонений фактических ординат точек от ординат должна быть наименьшей. Для парной линейной регрессии задача заключается в нахождении неизвестных параметров a и b, которые минимизируют сумму квадратов отклонений. Данные неизвестные параметры находятся из следующих уравнений:

$$a = \overline{y} - b\overline{x}$$

$$b = \frac{(\overline{xy} - \overline{x} - \overline{y})}{\frac{1}{n}\sum x_i^2 - \overline{x^2}}$$

Для оценки параметров парной регрессии необходимо, чтобы существовали следующие предпосылки:

- для генеральной совокупности связь между x и y должна быть линейной;
- наличие случайных отклонений, которые возникают при воздействии на переменную y множеством других, которые не учитываются в уравнении факторов и ошибок измерения.

При наличии таких предпосылок связь величин x_i и y_i приобретает вид: $y_i = \alpha + \beta x_i + \varepsilon_i$.

В данном уравнении ε_i - случайные ошибки. Из-за того, что значение отклонений ε_i неизвестны, то оценки параметров α и β, которые являются

случайными величинами, так как соответствуют случайной выборке, можно получить из наблюдений x_i и y_i.

Если a - оценка параметра α, а b - оценка параметра β, то оцененное уравнение регрессии приобретает вид: $y = bx_i + e_i$, где $_i$ – наблюдаемые значения ошибок ε_i.

При этом должны осуществляться предпосылки об отклонениях ε_i:

- величина ε_i является случайной переменной;
- математическое ожидание ε_i равна нулю;
- дисперсия ε_i постоянна: $D(\varepsilon_i) = D(\varepsilon_j)$ для всех i, j;
- значения ε_i независимы между собой.

Если данные условия выполняются, то оценки, полученные с помощью МНК, обладают следующими свойствами:

- Оценки являются несмещенными, т.е математическое ожидание оценки каждого параметра равно его истинному значению;
- Оценки состоятельны, т.к. дисперсия оценок параметров при возрастании числа наблюдений стремится к нулю.
- Оценки эффективны, они имеют наименьшую дисперсию по сравнению с другими оценками данного параметра.
- Оценка параметров определенной регрессии является только отдельным этапом длительного и сложного процесса построения эконометрической модели, так как первое оцененное уравнение редко удовлетворяет всем отношениям. Так же необходимо на каждом этапе анализировать качество оцененной зависимости. Данный анализ включает статистическую и содержательную составляющую.

Проверка статистического качества оцененного уравнения состоит из следующих этапов:

- проверка статистической значимости каждого коэффициента уравнения регрессии;
- проверка общего качества уравнения регрессии;
- проверка свойств данных, выполнение которых предполагалось при оценивании уравнения.

При содержательной составляющей анализа качества понимается рассмотрение экономического смысла оцененного уравнения регрессии.

При анализе общего качества оцененной линейной регрессии используется коэффициент детерминации R^2, называемый также квадратом множественной корреляции. Для случая парной регрессии это квадрат коэффициента корреляции переменных x и y. Коэффициент детерминации рассчитывается по следующей формуле:

$$R^2 = \frac{1 - \sum e_i^2}{\sum_i (y_i - \overline{y})^2}$$

Коэффициент детерминации является мерой, определяющей в какой степени найденная регрессионная прямая дает лучший результат для объяснения поведения зависимой переменной y, чем просто горизонтальная прямая $y = \overline{y}$.

Если существует статистически значимая линейная связь величин x и y, то R^2 будет ближе к единице. Однако, существуют случаи, что коэффициент детерминации близок к нулю в силу того, что обе эти величины имеют выраженный временной тренд, не связанный с их причинно-следственной зависимостью.

Для определения статистической значимости коэффициента детерминации проверяется нулевая гипотеза для F-статистики, которая находится по формуле:

$$F = \frac{R^2}{1 - R^2} * \frac{n - m - 1}{m}.$$

Смысл проверяемой гипотезы заключается в том, что все коэффициенты линейной регрессии, за исключением свободного параметра, равны нулю.

Распределение Фишера используется не только для проверки гипотезы об одновременном равенстве нулю всех коэффициентов линейной регрессии, но и гипотезы о равенстве нулю части этих коэффициентов. При анализе адекватности уравнения регрессии исследуемому процессу возможны следующие варианты:

- Построенная модель на основе ее проверки по F - критерию Фишера в целом адекватна, и все ее коэффициенты регрессии значимы. Данная модель подходит для принятия решений и составления прогнозов.

- Модель по F- критерию адекватна, но часть коэффициентов регрессии незначима. Данная модель подходит для принятия решение, но не подходит для составления прогнозов.
- Модель по F критерию адекватна, но все коэффициенты незначимы. Данная модель является полностью неадекватной, т.е. она не пригодна для принятия решений и составления прогнозов.

Анализ структуры временных рядов, содержащих сезонные или циклические колебания можно провести на основе расчета значений сезонной компоненты методом скользящей средней и построение аддитивной или мультипликативной модели временного ряда.

Аддитивная модель выглядит следующим образом: $Y = T + S + E$.

Мультипликативная модель выглядит так: $Y = T * S * E$, где T –трендовый компонент, S – сезонный компонент и E – случайный компонент. Выбор одной из двух моделей проводится на основе анализа структуры сезонных колебаний. Если амплитуда колебаний приблизительно постоянна, строят аддитивную модель временного ряда, если амплитуда сезонных колебаний возрастает или уменьшается, строят мультипликативную модель временного ряда, которая ставит уровни ряда в зависимость от значений сезонной компоненты.

Построение аддитивной и мультипликативной моделей сводится к расчету значений T, S и E для каждого уровня ряда.

Во **втором разделе** приводится описание подходов анализа сезонности электронной торговли.

Метод абсолютных и относительных разностей и индекс сезонности заключается в расчете сезонности относительно средних показателей. Для выделения сезонной волны определялся средний уровень $\overline{y_t}$ и общее среднее за данный период $\overline{y_c}$. Разница между этими показателями является характеристикой сезонности : $\Delta_{\text{сез}} = \overline{y_t} - \overline{y_c}$ в рамках метода абсолютных разностей, относительная сезонная неравномерность рассчитывается по формуле: $\Delta_{\text{отн}} = \frac{y_t - y_c}{y_c}$ и индекс сезонности определяется как: $I_{\text{сез}} = \frac{y_t}{y_c}$.

Корреляционно-регрессионный анализ и анализ сезонности заключается в выполнении следующих этапов:

1. Проверка значимости оцененного коэффициента регрессии в формуле $\widehat{y} = a + bt$ с помощью отношения к своему стандартному отклонению $t_b = \frac{b_\sigma}{b}$, $t_a = \frac{a_\sigma}{a}$.
2. Для анализа общего качества регрессии используют коэффициент детерминации R^2, значимость которого определяется на основе - статистики, имеющей для парной регрессии следующий вид :

$$F = \frac{R^2(n-2)}{(1-R^2)}$$

3. Процесс построения данной модели состоит из следующих этапов:
4. определение вида модели;
5. определение длины интервала сглаживания и использование метода скользящей средней для выравнивания исходного ряда;
6. определение значений сезонной компоненты и устранение ее влияния на исходные уровни ряда;
7. получение выровненных данных и аналитическое выравнивание уровней;
8. расчет полученных по модели значений, проверка качества полученной модели.

Прогноз по модели Хольта-Уинтерса на h шагов вперед определяется формулой:

$$Y_(t+h) = a_t + h * b_t t + s_{t-h+1+(h-1)modp},$$

где a_t, b_t и s_t рассчитываются следующим образом:

Расчет экспоненциально-сглаженного ряда:

$$a_t = \alpha(Y_t - s_t - p) + (1-\alpha)(a_{t-1} + b_{t-1})$$

Определение тренда:

$$b_t = \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1}$$

Оценка сезонности:

$$s_t = \gamma(Y_t - a_t) + (1 - \gamma)s_{t-p}$$

Коэффициенты сглаживания ряда α, тренда β и сезонности γ, лежат в диапазоне $[0, 1]$ и подгоняются таким образом, чтобы прогнозная кривая максимально точно повторяла исходные данные (критерий оценки – величина среднеквадратичной ошибки.

В **третьем разделе** были проведены все выше описанные методы анализа на основе торговых площадок Price и Яндекс.Маркет в 2013-2018 годах при помощи языка R, построены временные графики, графики по методу Хольта-Уинтерса, которые приведены в соответствии с рисунками 1 - 4 и сделан прогноз для 2019-2020 годов.

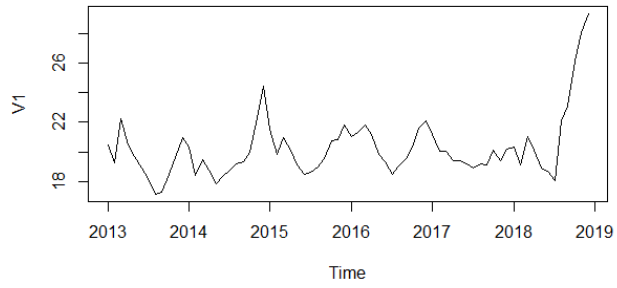

Рисунок 1 — Временной график для площадки Яндекс.Маркет

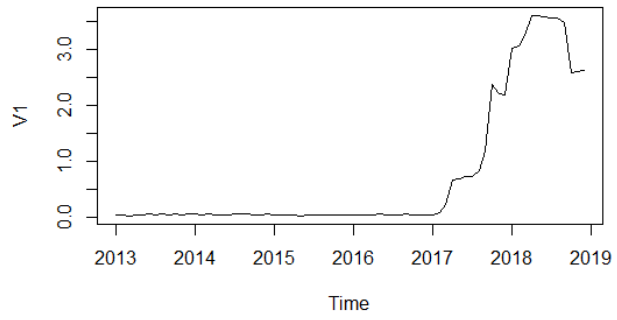

Рисунок 2 — Временной график для площадки Price

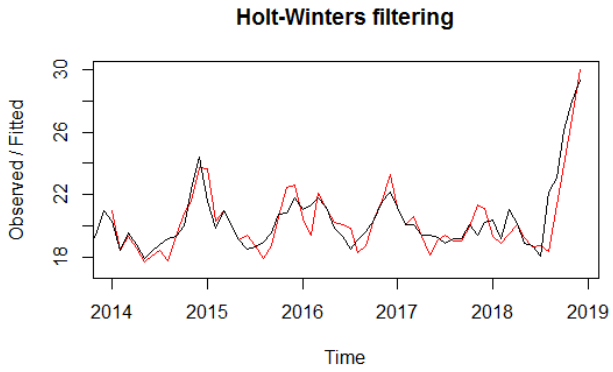

Рисунок 3 — График аппроксимации Хольта-Винтерса для площадки Яндекс.Маркет

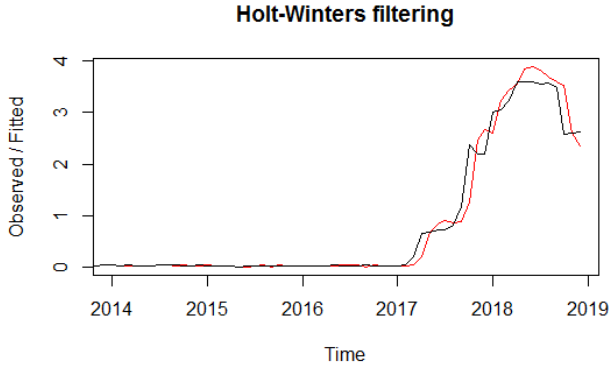

Рисунок 4 — График аппроксимации Хольта-Винтерса для площадки Price

В **заключении** бакалаврской работы приведены следующие результаты:

1. Определены основные понятия математических аспектов эконометрического анализа. А именно изучены метод наименьших квадратов, анализ статистической значимости параметров линейной модели, проверка общего качества уравнения линейной регрессии, F-тест на качество оценивания, моделирование сезонных и циклических колебаний для временного ряда.
2. Определены основные понятия, связанные с анализом сезонности электронной торговли.
3. Проведены анализ торговых площадок Яндекс.Маркет и Price в 2013-2018 годах и сделан прогноз для 2020-2021 годах.