УДК 004

Исследование зависимости рабочих мест в Лондоне и Великобритании с использованием метода линейной регрессии в системе R

Кочитов Михаил Евгеньевич
Приамурский государственный университет им. Шолом-Алейхема
студент

Баженов Руслан Иванович
Приамурский государственный университет им. Шолом-Алейхема
к.п.н., доцент, зав. кафедрой информационных систем, математики и правовой информатики

Аннотация
В данной статье исследуется зависимость рабочих мест в Лондоне и Великобритании с использованием интеллектуального метода линейной регрессии в системе R. Зависимость использованных данных является хорошей.
Ключевые слова: R, система, RStudio, регрессия, линейная, интеллектуальный метод, Лондон, Великобритания, рабочие места

The study of the dependence of jobs in London and the UK using the method of linear regression in the system R

Kochitov Mikhail Evgenevich
Sholom-Aleichem Priamursky State University
student

Bazhenov Ruslan Ivanovich
Sholom-Aleichem Priamursky State University
Candidate of pedagogical sciences, associate professor, Head of the Department of Information Systems, Mathematics and Law Informatics

Abstract
This article investigates the normal use of linear regression in the system R. The dependence of the data used is good.
Keywords: R, system, RStudio, regression, linear, intellectual method, London, UK, jobs

Большинство фирм, корпораций, учреждений, предприятий имеют огромное количество данных, которые каждый год меняются в лучшую или худшую сторону и чтобы спрогнозировать эти данные и узнать какими они будут в последующих годах, то используются интеллектуальные методы статистического анализа данных. В зависимости, какие данные собираемся

анализировать, то нам необходимо выбрать для этого эффективный интеллектуальный метод, который даст наиболее существенный и правдивый прогноз самих данных на будущий период и покажет в них зависимость, что лучше, а что хуже будет.

В статье В.В.Стрижкова и Р.А.Сологуба применили алгоритм выбора нелинейных регрессионных моделей [1]. И.Ю.Глухих разработал модель экспресс-анализа финансовой состоятельности организаций на базе методов многомерного регрессионного анализа [2]. В статье Л.М.Бугаевского и Г.Г.Прохорова рассматривается разработка методики составления карт взаимосвязи с использованием корреляционного и регрессионного анализов [3]. В.С.Баджанов и Е.А.Матушевская применили корреляционно-регрессионный анализ для анализа себестоимости продукции на примере ГУП АО "Севастопольский Винодельческий завод" [4]. В статье А.А.Манцаева проведен анализ долговременных тенденций производительности труда в РК: корреляционно-регрессионный анализ [5]. О. Ю.Легкодух, Д. Д. Капустина и Т.А.Кокодей провели анализ и прогноз динамики курса доллара, используя инструментарий регрессионного анализа [6]. В статье Н.Х.Рашитова был проведен анализ эффективности структуры экономики на основе корреляционно-регрессионного анализа [7]. R.Boukezzoula, S. Galichet и D.Coquin в своей статье рассматривали анализ параметрических интервальных регрессионных методологий в соответствии с онтологическими и эпистемическими видениями интервалов [8]. В статье F.O. de França исследуется математическое выражение, описывающее взаимосвязь набора объясняющих переменных с измеряемой переменной [9]. I. Georgiev и др. рассмотрели тесты на структурные изменения, основанные на статистических данных типа SupF и Cramer-von-Mises Andrews и Nyblom [10].

Целью данной статьи является исследование зависимости рабочих мест в Лондоне и Великобритании с использованием интеллектуального метода линейной регрессии в системе R.

В нашем случае данными будут являться рабочие места в Лондоне и Великобритании за период с 2005 по 2017 год. Эти данные были взяты с официального сайта Соединенного Королевства Британии и Северной Ирландии, иными словами Великобритании. Сами данные разделяются на три группы: общие, наемные (работа по найму) и самостоятельные (работа на себя) рабочие места в Лондоне и Великобритании, рабочие места в Лондоне и Великобритании по половому признаку и рабочие места Великобритании по регионам. Период времени данных разделен на 4 сезона (квартала) на каждый год, начиная с первого месяца каждого сезона, то есть март, июнь, сентябрь и декабрь [11].

Для анализа и выявление зависимостей данных рабочих мест используется интеллектуальный метод под названием «Метод линейной регрессии». Данный метод позволяет построить регрессионную модель, в которой можем увидеть зависимость этих данных за каждый год и как они менялись, а также построить саму линию регрессии, которая покажет, что с

данными будет происходить дальше, то есть какой будет вероятный их прогноз и в каком промежутке.

Метод линейной регрессии использует уравнение линейной регрессии y = a+bx (это уравнение похоже на уравнение прямой), где y – зависимая переменная, a – свободный член (пересечение) линии оценки, b – угловой коэффициент или градиент оценённой линии, x – независимая переменная [12].

Чтобы узнать, насколько хороша связь между двумя переменными, то вводится коэффициент детерминации (R^2). Чтобы вычислить этот коэффициент вручную, то потребуется множество формул. Первая формула это вычисления средних значений каждой переменной x и y и сразу обеих. Она вычисляется суммой всех значений деленных на их количество. Далее вычисляется дисперсия: сумма каждого значения возведенного в квадрат делится на их количество и вычитается средним значением, возведенным в квадрат. Потом вычисляется среднеквадратичное отклонение равное квадратному корню дисперсии. После вычисляется сам коэффициент корреляции в виде дроби: в числитель дроби идет разница между средним значением двух переменных x и y и умноженных средних значений каждой переменной x и y, и в знаменатель идет умножение среднеквадратичных отклонений. Ну и напоследок при найденном коэффициенте корреляции, можно уже найти сам коэффициент детерминации R^2 он равен всего лишь коэффициенту корреляции возведенному в квадрат [13].

Инструмент, который понадобится для реализации и использования метода линейной регрессии на данные, называется язык программирования R, который в основном предназначен для статистического анализа и обработки данных, а также в нем есть поддержка построение различных графиков и он является свободной средой, в которой можно проводить различные вычисления. Программное обеспечение, на котором понадобится использовать данный метод называется «RStudio», сама программа обрела популярность, благодаря ее понятно интуитивному интерфейсу и в ней гораздо удобнее работать, так как она имеет ряд инструментов, которые понадобятся для статистического анализа. В язык программирования R разработано было большое количество пакетов (библиотек), которые добавляют новые методы обработки данных, а также внедряют новые функции и возможности.

Теперь приступим к использованию самой программы «RStudio» и начнем использовать метод линейной регрессии на данные рабочих мест в Лондоне и Великобритании. Для начала выявим чему равен R^2 (коэффициент детерминации) используя функцию «lm» рабочих мест в первой группе.

```
Call:
lm(formula = Total.workforce.jobs.UK ~ Total.workforce.jobs.London,
    data = a)

Residuals:
    Min      1Q  Median      3Q     Max
-693062 -238964  -15071  270222  554693

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 1.737e+07  6.450e+05   26.93   <2e-16 ***
Total.workforce.jobs.London 2.961e+00  1.247e-01   23.74   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 344900 on 50 degrees of freedom
Multiple R-squared:  0.9185,    Adjusted R-squared:  0.9169
F-statistic: 563.8 on 1 and 50 DF,  p-value: < 2.2e-16
```

Рис 1. Результат функции «lm»

Как видим на рисунке 1, показан результат использования функции «lm» зависимых данных общих рабочих мест в Великобритании и Лондоне, R^2 = 0.9169, это означает, что если R^2 приближен к 0.8 или выше его, то линейная регрессия этих зависимых данных имеет более хорошую связь.

К оставшимся данным первой группы были выданы результаты R^2 ниже:

- Наемные рабочие места в Великобритании и Лондоне = 0.8146;
- Самостоятельные рабочие места в Великобритании и Лондоне = 0.9359.

Далее необходимо наглядно посмотреть зависимости данных первой группы с методом линейной регрессии, для этого необходимо построить график.

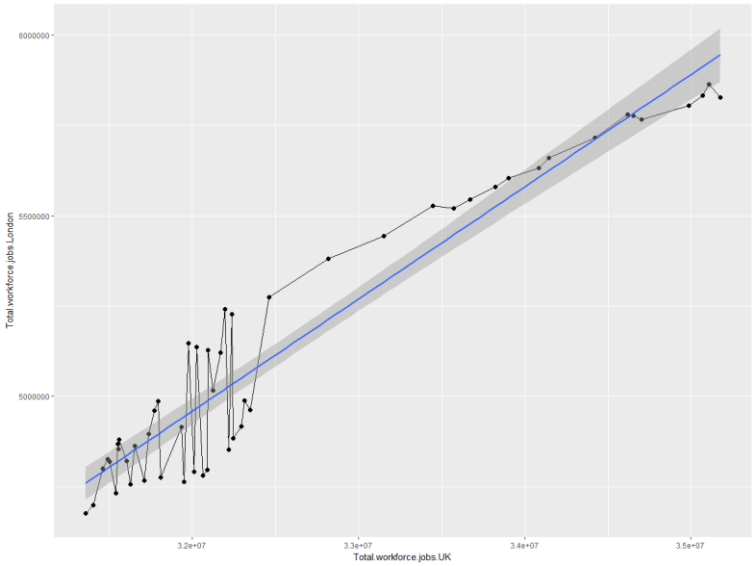

Рис. 2. График зависимости общих рабочих мест в Великобритании и Лондоне с отображением линии тренда

На рисунке 2 изображен график, в нем видим зависимость двух данных: общих рабочих мест в Великобритании и Лондоне. Точки на графике являются кварталами (сезонами) годов с 2005 по 2017, то есть каждые 4 точки это следующий год. Как видно на рисунке в начале периода с 2005 по 2010 были скачки, так как в этих годах количество общих рабочих мест в Лондоне и Великобритании менялось с уменьшением и увеличением.

После примерно 2010 года, количество общих рабочих мест начало постепенно расти, что показывает линейная регрессия, изображенная линией тренда, которая стремится вверх.

Далее рассмотрим график зависимости данных наемных рабочих мест в Лондоне и Великобритании.

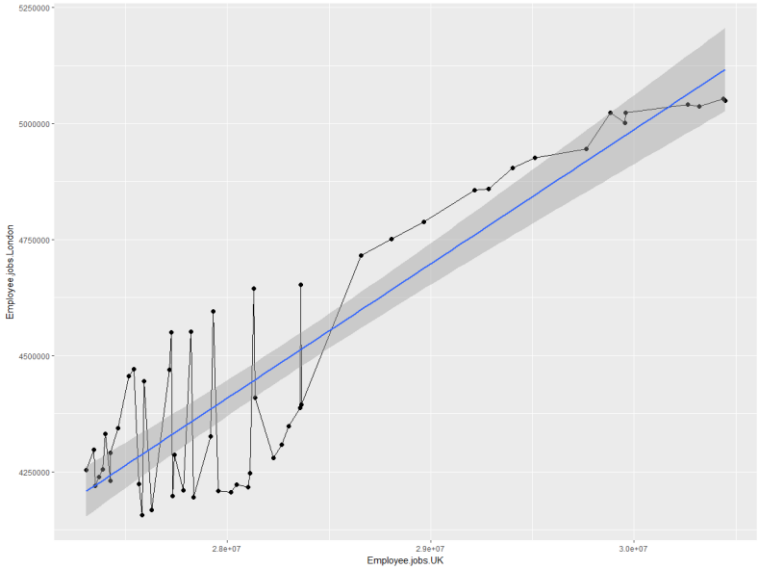

Рис. 3. График зависимости наемных рабочих мест в Великобритании и Лондоне с отображением линии тренда

На рисунке 3 график немного схож с графиком, изображенным на рисунке 2. В нем также примерно до 2010 года происходили скачки с сокращением и увеличением количества наемных рабочих мест в Лондоне и Великобритании, а после 2010 года, наемные рабочие места стали постепенно расти, чему свидетельствует также линия тренда, которая растет вверх.

Аналогично построим график самостоятельных рабочих мест в Лондоне и Великобритании.

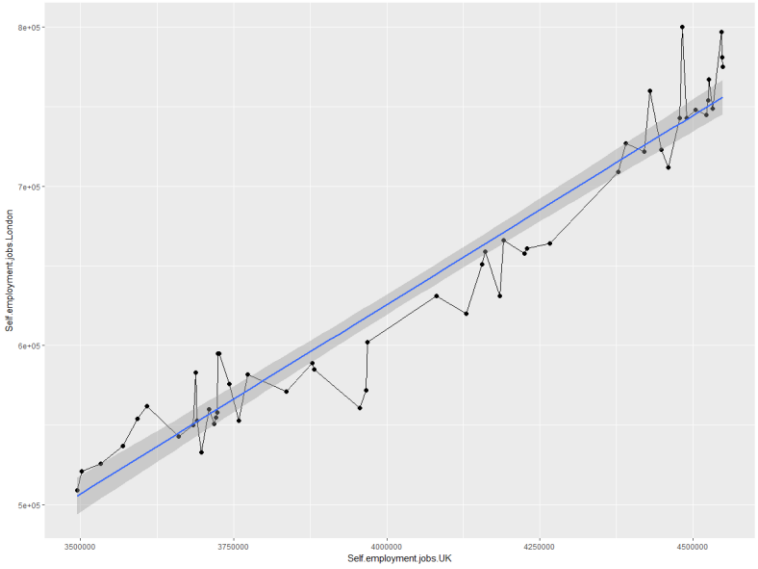

Рис. 4. График зависимости самостоятельных рабочих мест в Великобритании и Лондоне с отображением линии тренда

На рисунке 4 график немного отличается тем, что зависимость данных самостоятельных рабочих мест в Лондоне и Великобритании меняется с каждым сезоном, иными словами количество этих рабочих мест, может быть меньше или больше в зависимости от сезона. Однако все же линейная регрессия показывает, что данные постепенно растут с множественными падениями, то можно предположить, что в будущем, количество самостоятельных рабочих мест будет значительно повышаться скачками.

Теперь перейдем к данным второй группы, а это рабочие места в Великобритании и Лондоне по половому признаку. Для начала используем функцию «lm» для нахождения R^2. Все его значения представлены ниже:

- Рабочие места, занятые мужчинами в Великобритании и Лондоне = 0.9351;
- Рабочие места, занятые женщинами в Великобритании и Лондоне = 0.8792;
- Рабочие места, занятые мужчинами и женщинами в Великобритании = 0.957;
- Рабочие места, занятые мужчинами и женщинами в Лондоне = 0.9326.

Как видно значения R^2 все выше 0.8, это значит, что зависимости этих данных имеют более хорошую связь.

Теперь построим собственно сами графики (Рис. 5, 6, 7, 8).

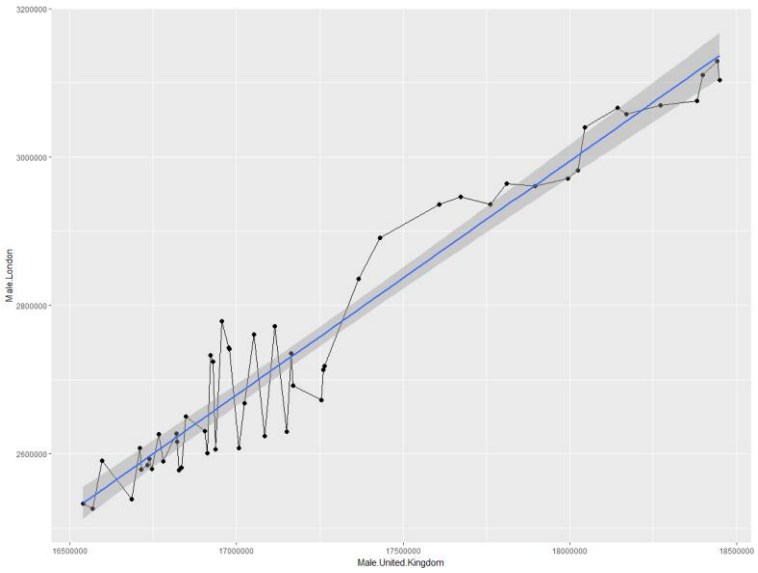

Рис. 5. График зависимости рабочих мест, занятых мужчинами в Великобритании и Лондоне с отображением линии тренда

На рисунке 5 аналогично показано, что до 2010 года были скачки, а после 2010 года, количество рабочих мест, занятых мужчинами в Великобритании и Лондоне стало постепенно расти.

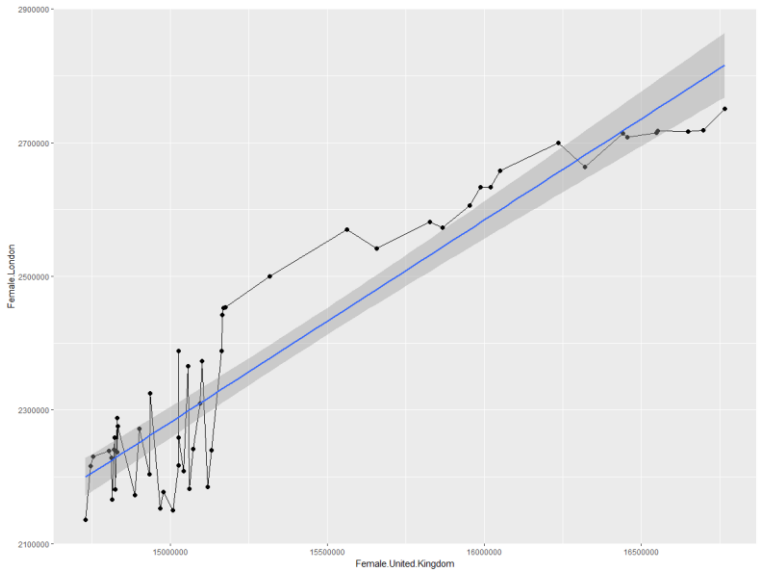

Рис. 6. График зависимости рабочих мест, занятых женщинами в Великобритании и Лондоне с отображением линии тренда

На рисунке 6 также продемонстрирована схожесть, что и с рисунком 5, однако, количество рабочих мест, занятых женщинами начало постепенно расти уже ранее 2010 года, чем рабочие места, занятых мужчинами.

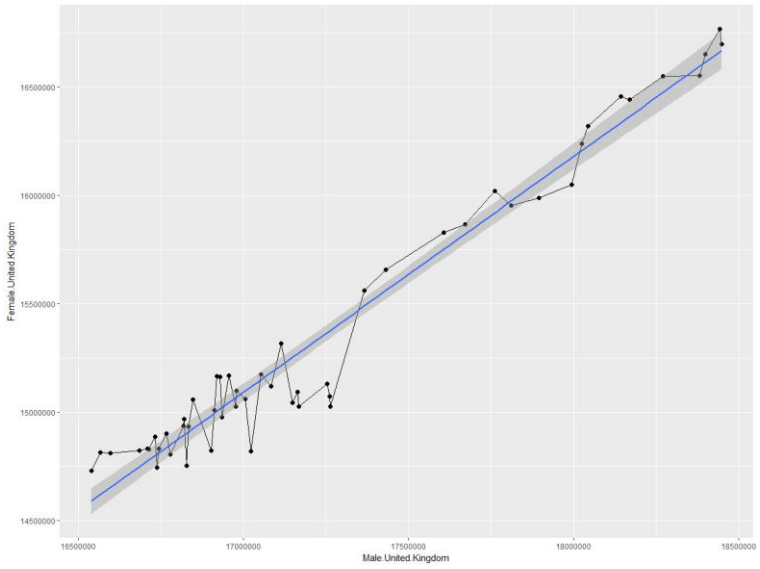

Рис. 7. График зависимости рабочих мест, занятых мужчинами и женщинами в Великобритании с отображением линии тренда

На рисунке 7 представлен график, в котором изменение почти схожи с предыдущими графиками, так как в различные сезоны начальных годов (с 2005 по 2008) происходили скачки, дальше пошел постепенный рост количества рабочих мест, занятых мужчинами и женщинами в Великобритании, что прогнозирует сами линия тренда.

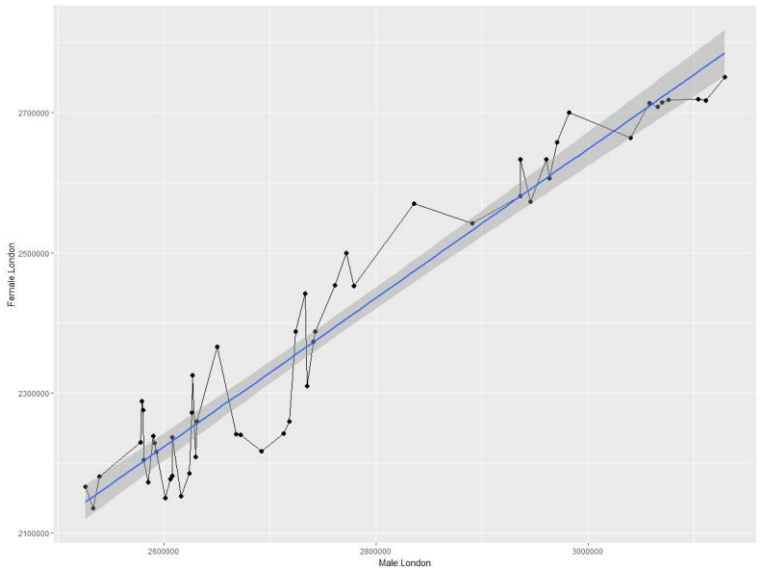

Рис. 8. График зависимости рабочих мест, занятых мужчинами и женщинами в Лондоне с отображением линии тренда

График, представленный на рисунке 8, показывает зависимость рабочих мест, занятых мужчинами и женщинами в Лондоне. Здесь можно заметить, что примерно с 2009 по 2010 год, произошел резкое сокращение рабочих мест и медленный их рост в течение этого периода. Далее рабочие места, занятые мужчинами и женщинами то падали с каждым годом, то росли, однако они все же постепенно в будущем будут расти, что свидетельствует линия регрессии.

Теперь осталось провести линейную регрессию на данных третьей группы: рабочие места в Великобритании по регионам. Сама Великобритания делится на четыре главных региона: Англия (Южная, восточная и центральная часть Великобритании, в ней находится столица Лондон), Уэльс (западная часть Великобритании), Шотландия (северная часть Великобритании) и Северная Ирландия (северная часть острова Ирландии). Данные третьей группы имеют наиболее большой период с 1996 по 2017 год.

Используем функцию «lm» и смотрим значения R^2:

- Общие рабочие места в Англии и Шотландии = 0.7375;
- Общие рабочие места в Уэльс и Северной Ирландии = 0.8769;
- Общие рабочие места в Уэльс и Англии = 0.9164;
- Общие рабочие места в Шотландии и Северной Ирландии = 0.8652.

Как можно заметить, что по зависимости данных общих рабочих мест в Англии и Шотландии R^2 меньше 0.8, однако при округлении он все же идет к 0.8, это означает, что такая зависимость имеет неплохую связь, а другие намного лучше.

Теперь осталось построить сами графики зависимостей (Рис. 9, 10, 11, 12).

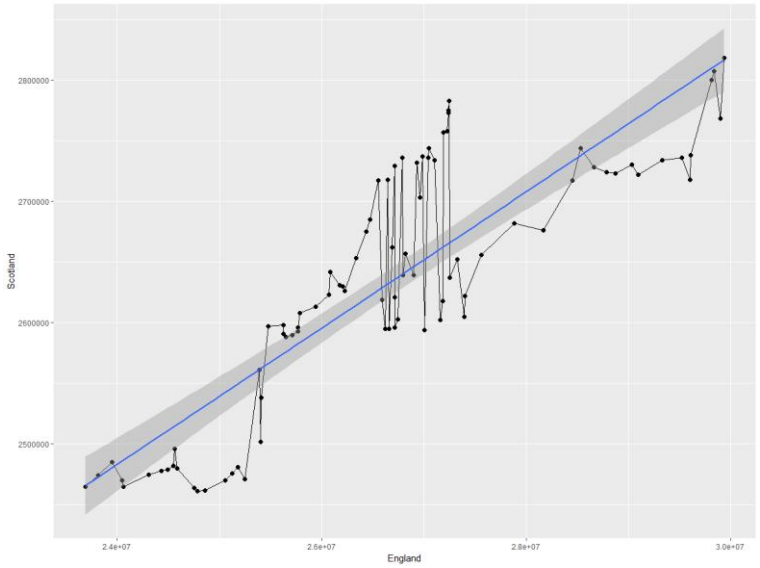

Рис. 9. График зависимости общих рабочих мест в Англии и Шотландии с отображением линии тренда

График, изображенный на рисунке 9, с довольно большим периодом с 1996 по 2017 год имеет, уже другую зависимость, как можно заметить ближе к 2000 годам, произошел резкий рост рабочих мест в Англии и Шотландии. Однако где то с 2006 по 2010 год по сезонам были значительные резкие скачки с резким сокращением и увеличением рабочих мест в Англии и Шотландии. После этого периода, количество рабочих мест стало значительно расти с малейшими скачками.

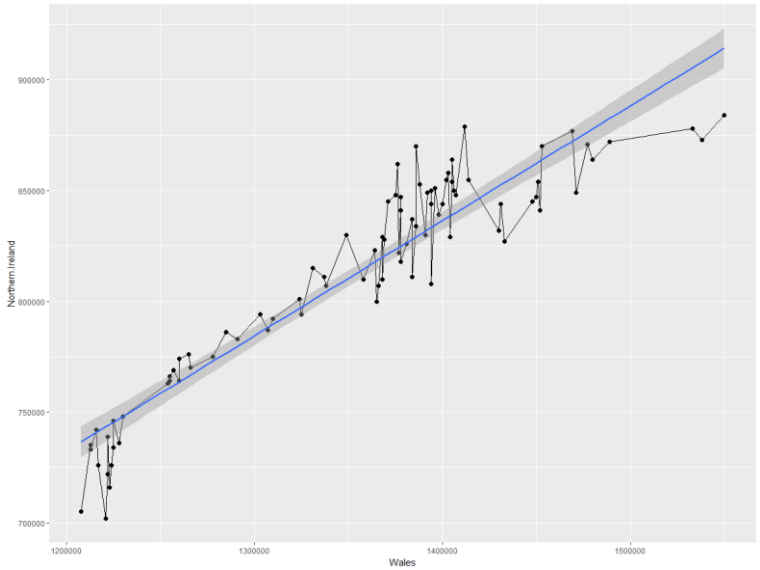

Рис. 10. График зависимости общих рабочих мест в Уэльс и Северной Ирландии с отображением линии тренда

На рисунке 10 представлен график, в котором зависимость, связанные с количеством рабочих мест в Уэльс и Северной Ирландии в периоде с 1996 по 1998 год имела скачки. После этого пошел постепенный рост, но с 2006 по 2010 также как и в предыдущем графике были постепенные скачки по

сезонам. Минуя эти периоды, пошел слегка рост рабочих мест в Уэльс и Северной Ирландии.

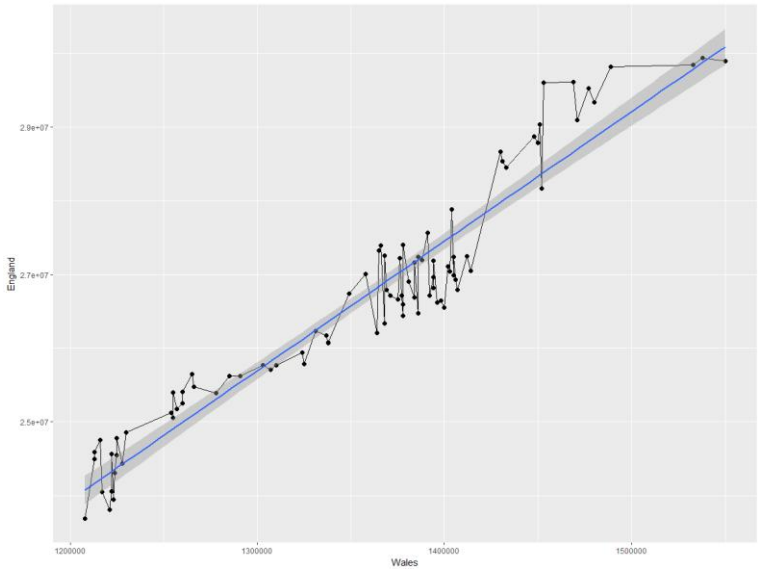

Рис. 11. График зависимости общих рабочих мест в Уэльс и Англии с отображением линии тренда

Рассматривая график на рисунке 11 можно предположить то же самое, что и на графике, изображенном на рисунке 10. В нем также скачки с 1996 по 1998 и с 2006 по 2010 год, но в 2011 года произошел резкий рост количества рабочих мест в Уэльс и Англии и стал на протяжении следующих годов стабильно держаться с незначительными спадами.

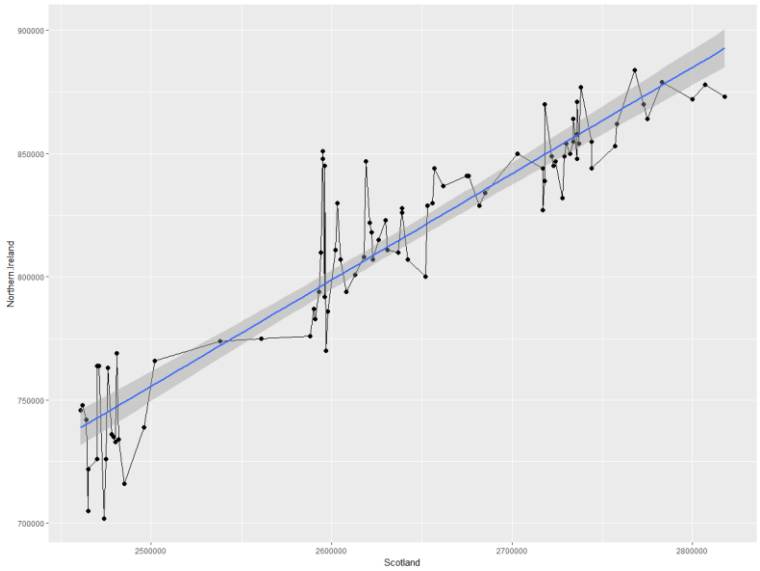

Рис. 12. График зависимости общих рабочих мест в Шотландии и Северной Ирландии с отображением линии тренда

Глядя на последний график на рисунке 12, можно заметить, что также с 1996 по 1998 год были скачки, потом резкий рост. Далее с 2006 по 2010 год, были также очень большие скачки, но еще с 2013 по 2015 были еще

несколько скачков в количестве рабочих мест в Шотландии и Северной Ирландии, однако дальше пошел постепенный рост.

Таким образом, в данной статье была достигнута цель с исследованием зависимостей данных о рабочих местах в Великобритании и Лондоне с использованием метода линейной регрессии в системе R. Можно, судя по графикам, можно заметить, что в начале, были чередующиеся сокращения и увеличения рабочих мест в Великобритании и Лондоне в период с 2007 по 2010 год, однако это прекратилось незначительно и с каждым годом стали данные постепенно расти с небольшими спадами. Прогноз линейной регрессии дает нам вывод, что возможно в будущем рабочие места будут также расти и возможно в какой-то период произойдет возможный скачок.

Библиографический список

1. Стрижов В.В., Сологуб Р.А. Алгоритм выбора нелинейных регрессионных моделей с анализом гиперпараметров // Математические методы распознавания образов. 2009. Т. 14. № 1. С. 184-187.
2. Глухих И.Ю. Разработка моделей экспресс-анализа финансовой состоятельности организаций на базе методов многомерного регрессионного анализа // Управленческое консультирование. Актуальные проблемы государственного и муниципального управления. 2011. № 3 (43). С. 185-195.
3. Бугаевский Л.М., Прохоров Г.Г. Разработка методики составления карт взаимосвязи с использованием корреляционного и регрессионного анализов // Известия высших учебных заведений. Геодезия и аэрофотосъемка. 1990. № 6. С. 101-109.
4. Баджанов В.С., Матушевская Е.А. Применение корреляционно-регрессионного анализа для анализа себестоимости продукции на примере ГУП АО "Севастопольский Винодельческий завод" // Southern Almanac of Scientific Research. 2017. № 4 (4). С. 20-25.
5. Манцаева А.А. Анализ долговременных тенденций производительности труда в рк: корреляционно-регрессионный анализ // Вестник Института комплексных исследований аридных территорий. 2015. Т. 1. № 1 (30). С. 18-24.
6. Легкодух О.Ю., Капустина Д.Д., Кокодей Т.А. Анализ и прогноз динамики курса доллара, используя инструментарий регрессионного анализа // В сборнике: Развитие методологии современной экономической науки и менеджмента материалы I Всероссийской конференции студентов, аспирантов и молодых учёных. Севастопольский государственный университет. 2016. С. 57-58.
7. Рашитова Н.Х. Анализ эффективности структуры экономики на основе корреляционно-регрессионного анализа // В сборнике: Инновационное развитие российской экономики материалы X Международной научно-практической конференции. Российской Федерации Российский экономический университет имени Г. В. Плеханова; Российский фонд

фундаментальных исследований. 2017. С. 250-252.
8. Boukezzoula R., Galichet S., Coquin D. From fuzzy regression to gradual regression: Interval-based analysis and extensions // Information Sciences. 2018. Т. 441. С. 18-40
9. França de F. O. A greedy search tree heuristic for symbolic regression // Information Sciences. 2018. Т. 442–443. С. 18-32
10. Georgiev I., Harvey D. I., Leybourne S.J., Taylor A.M.R. Testing for parameter instability in predictive regression models // Journal of Econometrics. 2018. Т.204. №1. С. 101-118
11. База данных в Лондоне. URL: https://data.london.gov.uk/dataset (дата обращения 14.04.2018)
12. Основы линейной регрессии. URL: http://statistica.ru/theory/osnovy-lineynoy-regressii (дата обращения 28.04.2018)
13. Пример нахождения коэффициента детерминации. URL: https://math.semestr.ru/corel/prim2.php (дата обращения 28.04.2018)