УДК:004:005

О НЕКОТОРЫХ «ПОДВОДНЫХ КАМНЯХ» ПРИ ИСПОЛЬЗОВАНИИ ПРОСТОЙ ЛИНЕЙНОЙ РЕГРЕССИИ И УНИВЕРСАЛЬНЫХ КОМПЬЮТЕРНЫХ СРЕДСТВ

ҚАРАПАЙЫМ СЫЗЫҚТЫҚ РЕГРЕССИЯНЫ ЖӘНЕ ӘМБЕБАП КОМПЬЮТЕРЛІК ҚҰРАЛДАРДЫ ҚОЛДАНУДА ТУЫНДАЙТЫН КЕЙБІР МӘСЕЛЕЛЕР ТУРАЛЫ

ON SOME "PITFALLS" WHEN USING SIMPLE LINEAR REGRESSION AND VERSATILE COMPUTER

В.З. КРУЧЕНЕЦКИЙ, А.А. КАЛАБИНА, В.В. КРУЧЕНЕЦКИЙ, А.Б. МИМЕНБАЕВА., Ж.К. СЕРИКУЛОВА, А.Р. ПАК
V.Z. KRUCHENETSKY, A.A. KALABINA, V.V. KRUCHENETSKY, A.B. MIMENBAEVA, J.K. SERIKULOVA, A.R. PAK

(Алматы технологиялық университеті)
(Алматинский технологический университет)
(Almaty Technological University)
E-mail: anesti-an@mail.ru

Рассматривается использование стандартных компьютерных средств, в частности табличного процессора MS Excel и надстройки к нему PH Stat2, для решения экономических задач, бизнес-процессов на основе применения простой линейной регрессии, доступных широкому кругу пользователей, проблемы, возникающие при этом, пути их преодоления.

Қарапайым сызықтық регрессияны қолдану негізінде экономикалық есептерді, бизнес-үрдістерді шешу үшін стандартты компьютерлік құралдарды, соның ішінде қолданушылардың кең ауқымына қол жетімді MS Excel кестелік процессорын және оның PH Stat2 қондырмасын қолдану барысында туындайтын мәселелерді шешу жолдары қарастырылады.

We consider the use of standard computer tools, including MS Excel spreadsheet and add to it PH Stat2, to address the economic problems of business processes through the use of simple linear regression, the available wide range of users, the problems arising in this case, how to overcome them.

Ключевые слова: аппроксимация, бизнес-процесс, гомоскедастичность, диаграмма разброса, корреляция, регрессия, тренд.

Негізгі сөздер: аппроксимация, бизнес-үрдіс, гомоскедастілік, шашырау диаграммасы, корреляция, регрессия, тренд.

Keywords: approximation, business process, homoscedasticity, scatter diagram, correlation, regression, trend.

Введение

Решение экономических задач, бизнес и многих других процессов без использования компьютерных средств сегодня трудно вообразить, ибо в силу их сложности и большой трудоемкости результаты могут носить приближенный характер, а нередко оказывается невозможными.

Компьютерные средства для решения указанных выше задач отличаются многообразием – это и универсальные пакеты прикладных программ, специализированные пакеты и комплексы программ, отдельные уникальные программы. Их использование, как правило, связано с необходимостью специальной подготовки пользователей, достаточно большими материальными затратами и трудоемкостью, практически всегда имеющими место ограничениями.

В данной работе рассматривается использование стандартных компьютерных средств для решения экономических задач и бизнес-процессов, доступных широкому кругу пользователей, таких, например, как

электронный процессор Microsoft Excel и достаточно мощная надстройка к этому процессору – PH Stat.

Объекты и методы исследований

Программный продукт MS Excel или, как это принято называть в экономике – программа калькуляции таблиц и деловой графики [1-3], удовлетворяет самым высоким запросам пользователей, постоянно совершенствуется и, даже, начиная с ранних версий 3.0÷6.0, не говоря о современных – 2000 ÷ 2010, является мощным универсальным средством, позволяющим решать разнообразные задачи: от расчета подоходного налога, до составления финансового отчета крупной корпорации. Возможности Excel не ограничиваются только выполнением вычислительных операций, они значительно шире – это и обработка текста, управление базами данных, что во многих случаях превосходит специализированные программы-редакторы или программы управления базами данных.

В MS Excel встроено множество функций. Он предлагает удобные средства создания простых и сложных формул, используя ссылки на ячейки, операторы и функции. Мастер функций позволяет легко строить сложнейшие формулы, которые можно редактировать, копировать и перемещать по рабочему листу и рабочей книге. Привлекательные черты этого стандартного программного продукта:

- является неотъемлемой частью рабочего места пользователей, поэтому отпадают затраты на дополнительное программное обеспечение;
- простота, как для обучения, так и для использования;
- графические и статистические функции, вычислительные и аналитические возможности MS Excel оперируют с теми же рабочими листами, которые пользователи применяют для хранения данных;
- часть графических функций процессора создает лучшую, более ясную визуализацию и представление данных, чем другие пакеты программ.
- MS Excel совместим со всеми известными приложениями Microsoft Office.

Сведения по использованию MS Excel весьма подробно описаны в многочисленных источниках; адресованы они как начинающим, так и профессиональным пользователям. В данной статье основное внимание обращено не на основы работы с MS Excel, а на использование в решении задач, связанных с моделированием бизнес-процессов. Что касается надстройки PH Stat, то она оказывается весьма удобной и эффективной, поскольку большинство задач бизнес-процессов, являются стохастическими, связанными с оптимизацией, прогнозированием и, следовательно, с использованием математических моделей, способов, приемов и методов математической статистики.

Результаты и их обсуждение

Исследование моделей, так называемой, коммерческой статистики в решении задач бизнес-процессов, имеет широкое распространение и возможности и представляет значительный интерес. Статистические методы применяют в самых разнообразных сферах бизнеса. В бухгалтерском учете они используются для извлечения и анализа выборок данных, подвергающихся аудиторской проверке, а также для определения затрат при исчислении себестоимости. В финансовом деле статистика позволяет принять правильное решение при выборе объектов капиталовложений и отслеживать финансовые показатели, изменяющиеся во времени. Менеджеры используют статистические методы для улучшения качества производимой продукции или предоставляемых услуг. В маркетинге статистика позволяет оценить долю клиентов, предпочитающих один вид продукции другому, выяснить причины этого явления, а также определить, какая из рекламных стратегий увеличивает сбыт продукции. То есть круг пользователей статистических методов весьма широкий.

Используя программу MS Excel, пользователь должен не только делать правильный выбор метода, но и хорошо знать условия его применения, иметь глубокое понимание математических моделей, статистических понятий, связанных с решаемой задачей, чтобы предотвратить некорректный анализ или другую ошибку. Одновременно для правильного применения MS Excel необходимо знать ограничения, которые налагаются, учитывать возможности, недостатки, памятуя о том, что не существует единого оптимального, абсолютного способа, процедур применения программы Microsoft Excel, который подошел бы абсолютно всем пользователям в решении многообразных задач бизнеса.

Поскольку в бизнес–процессах, решении экономических задач значительное место

занимают статистические данные, способы, методы и приемы их обработки, анализа, предсказания значений зависимой переменной, а также, учитывая ограниченный объем данной статьи, то основное внимание обращено на использование модели простой линейной регрессии, условия ее применимости и способы проверки этих условий.

Известно, что для определения тесноты связи между показателями, не находящимися в функциональной зависимости, широко используются методы корреляционного и регрессионного анализа, Корреляция представляет собой вероятностную зависимость между явлениями, не имеющую строго функционального характера. Корреляционная зависимость может быть выявлена как между двумя количественными признаками, так и между многими величинами. В последнем случае приходится иметь дело с множественной корреляцией.

Используя уравнение регрессии, можно получить удовлетворительное значение результативного признака только в том случае, если значения факторных признаков, подставляемых в уравнение регрессии, близки к тем эмпирическим значениям, на основе которых определяются параметры уравнения.

Непременным условием применения корреляционного и регрессионного анализа является обеспечение: репрезентативности статистических данных, обоснованность применения вероятностной схемы к изучаемому явлению. Практически это сводится к выбору уравнения соответствующей кривой – логарифмической, параболы, гиперболы и др.

Теснота связи между изучаемыми явлениями измеряется корреляционным отношением для криволинейной зависимости; для прямолинейной зависимости исчисляется коэффициент корреляции.

Регрессионные модели отличаются большим разнообразием. Зависимость между двумя переменными может быть разной: от самой простой до крайне сложной. Пример простейшей (линейной) зависимости показан на рис.1.

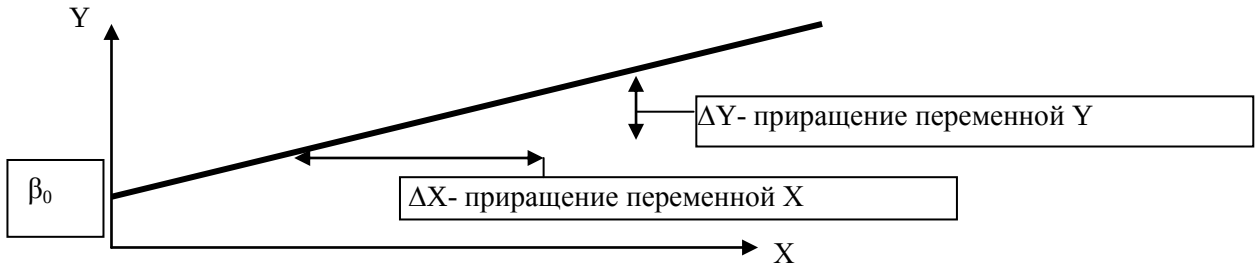

Рисунок 1 - Положительная линейная зависимость.

Простая линейная регрессия выражается как:

$$Y_i = \beta_0 + \beta_1 * x_i + \varepsilon_i \quad (1),$$

где: β_0 - сдвиг (длина отрезка, отсекаемого на координатной оси прямой Y), β_1 - наклон прямой Y, ε_i - случайная ошибка переменной Y в i - м наблюдении.

В этой модели наклон β_1 представляет собой количество единиц измерения переменной Y, приходящихся на одну единицу измерения переменной X. Эта величина характеризует среднюю величину измерения переменной Y(положительного или отрицательного) на заданном отрезке оси X. Сдвиг β_0 - представляет собой среднее значение переменной Y, когда переменная X равна 0. Последний компонент модели ε_i является случайной ошибкой переменной Y в i - м наблюдении.

Выбор подходящей математической модели зависит от распределения значений переменных X и Y на диаграмме разброса. Различные виды зависимости переменных показаны на рис. 2 а-е.

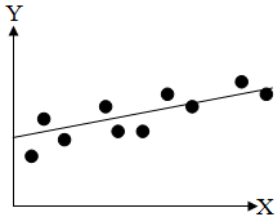

Рис. 2а Положительная линейная зависимость

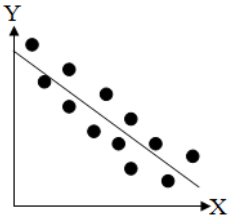

Рис. 2б Отрицательная линейная зависимость

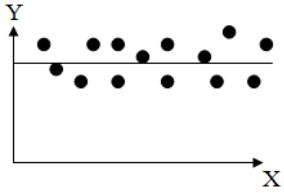

Рис. 2в Переменные X и Y не зависят друг от друга

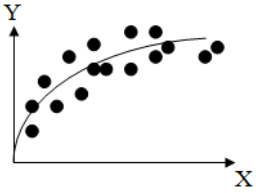

Рис. 2г Положительная криволинейная зависимость

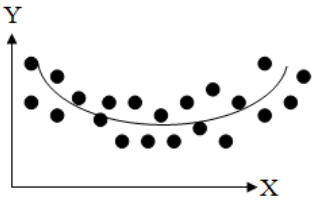

Рис. 2д Ụ - образная криволинейная зависимость

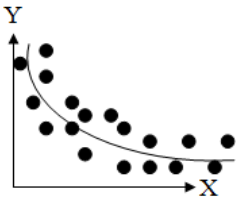

Рис. 2е Отрицательная криволинейная зависимость

Известно [2], что условиями применения регрессионного анализа являются следующие:

1. Ошибка должна иметь нормальное распределение;

2. Вариация данных вокруг линии регрессии должна быть постоянной (свойство гомоскедастичности);

3. Ошибки должны быть независимыми.

Первое предположение о нормальном распределении ошибок требует, чтобы при каждом значении переменной X ошибки линейной регрессии имели нормальное распределение. Второе условие – гомоскедастичность, заключается в том, что вариация данных вокруг линии регрессии должна быть постоянной при любом значении переменной X. Это означает, что величина ошибки, как при малых, так и при больших значениях переменной X должна изменяться в одном и том же интервале. Свойство гомоскедастичности очень важно для метода наименьших квадратов, с помощью которого определяются коэффициенты регрессии. Если это условие нарушается, то необходимо применять либо преобразование данных, либо метод наименьших квадратов с весами [2].

Третье предположение, о независимости ошибок, заключается в том, что ошибки регрессии не должны зависеть от значения переменной X. Это условие особенно важно, если данные собираются на протяжении определенного отрезка времени. В этих ситуациях ошибки, присущие конкретному отрезку времени, часто коррелируют с ошибками, характерными для предыдущего периода.

С учетом этих предположений в качестве оценки генеральной совокупности β_0 и β_1 можно использовать сдвиг b_0 и наклон b_1 прямой Y. Таким образом, уравнение регрессии принимает вид:

$$\acute{Y} = b_0 + b_1 * X_i \qquad (2),$$

где: $\acute{Y}$ – предсказанное значение переменной Y для i-го наблюдения; X_i - значение переменной X в i-м наблюдении.

Т.е., простая линейная регрессия выражается, как «предсказанное значение переменной Y, равное сумме сдвига и наклона, умноженное на значение переменной X».

Для того, чтобы предсказать значение переменной Y, в уравнении (2) необходимо определить два коэффициента регрессии – сдвиг b_0 и наклон b_1 прямой Y. Вычислив эти параметры, проведем прямую на диаграмме разброса. Затем можно визуально оценить, насколько близка регрессионная прямая к точкам наблюдения.

Критерии соответствия можно задать разными способами. Проще всего минимизи-

ровать разности между фактическими значениями Y и предсказанными значениями Ý. Однако, поскольку эти разности могут быть как положительными, так и отрицательными, следует минимизировать сумму их квадратов. Из уравнения (2) следует:

$$\sum_{i=1}^{n}(Y_i - \acute{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - (b_0 + b_1 * X_i))^2 \quad (3)$$

Здесь параметры b_0 и b_1 неизвестны. Таким образом, сумма квадратов разностей является функцией, зависящей от сдвига b_0 и наклона b_1 выборки Y. Для того, чтобы найти значения параметров b_0 и b_1, минимизирующих сумму квадратов разностей, применяется метод наименьших квадратов. Для иллюстрации данного положения построим диаграмму разброса на примере зависимости доходов сети магазинов от их площадей (статистика реальных данных для конкретной организации по годам взята за 12 лет).

Ниже (рис. 3) приводится диаграмма разброса, построенная с помощью MS EXCEL и нанесенной на нее прямой Y, т.е. линией регрессии, и найденным соот-ветственно ее уравнением:

$$Y = 0.9645 + 1.6699X \quad (4)$$

Как следует из указанной диаграммы, точность определения уравнения регрессии (аппроксимации) составляет 0.9042 (R_2= 0.9042), сдвиг b_0 - 0.9645, наклон b_1 - 1.6699.

Применяя регрессионную модель для прогнозирования, необходимо учитывать лишь допустимые значения независимой переменной. В этот диапазон входят все значения переменной X, начиная с минимальной и заканчивая максимальной. Таким образом, предсказывая значение переменной Y при конкретном значении переменной X, мы выполняем интерполяцию между значениями переменной X в диапазоне возможных значений. Однако экстраполяция значений за пределы этого интервала невозможна [2]. Любая попытка экстраполяции означает, что мы предполагаем, будто линейная регрессия сохраняет свой характер за пределами допустимого диапазона.

Для того чтобы предсказать значение зависимой переменной по значениям независимой переменной в рамках избранной статистической модели, необходимо оценить изменчивость. Существует несколько способов оценки изменчивости. Один из них использует общую сумму квадратов (total sum of squares – SST), позволяющую оценить колебания значений Y вокруг среднего значения $Y_{ср}$. Полная вариация, представляющая собой полную сумму квадратов, делится на объяснимую вариацию, или сумму квадратов регрессии (regression sum of squares – SSR, or explained variation) и необъяснимую вариацию (unexplained variation), или сумму квадратов ошибок (error sum of squares – SSE). Объяснимая вариация характеризует взаимосвязь между переменными X и Y, а необъяснимая зависит от других факторов (рис.4).

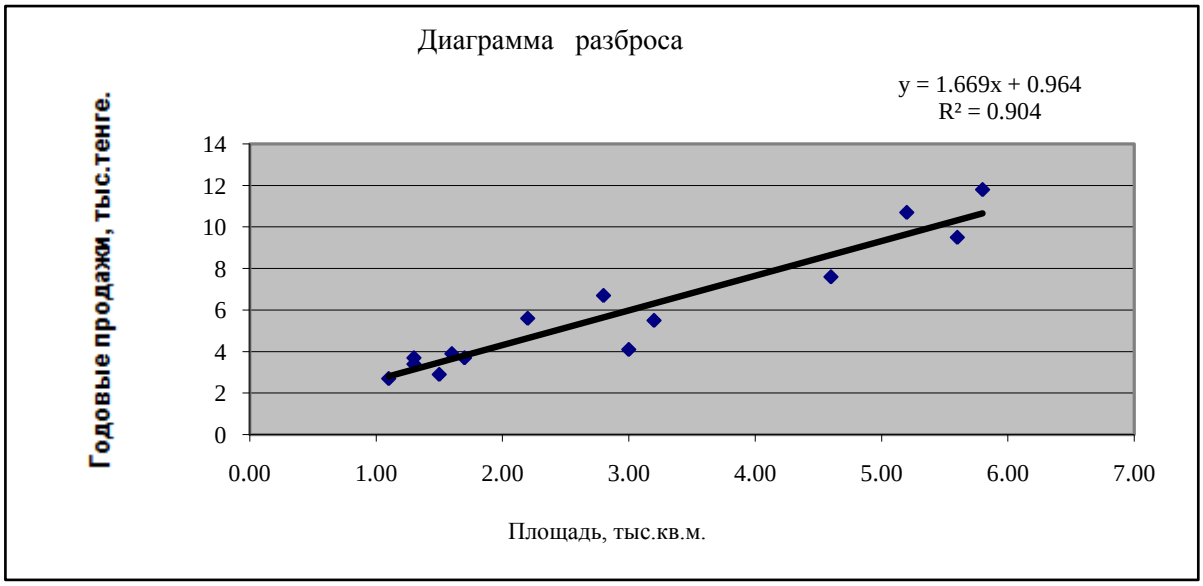

Рисунок 3 - Диаграмма разброса и линия, построенная с помощью программы MS EXCEL.

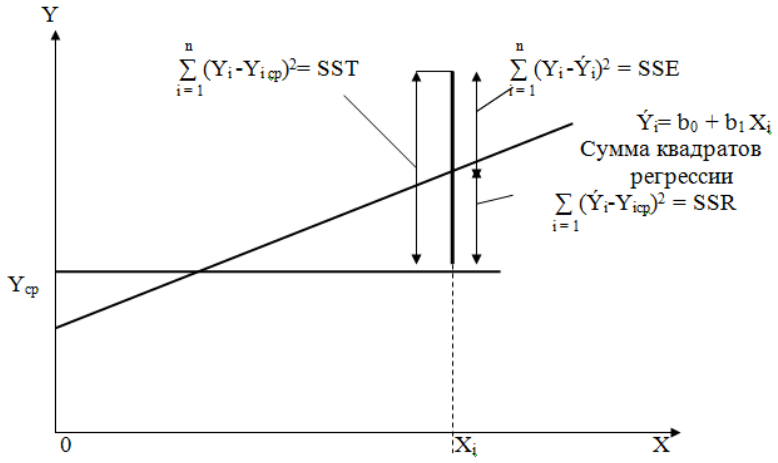

Рисунок 4 - Оценки изменчивости в модели регрессии.

Сумма квадратов регрессии (SSR) представляет собой сумму квадратов разностей между $\acute{Y}_i$ (предсказанным значением переменной Y) и $Y_{ср}$ (средним значением переменной Y). Сумма квадратов ошибок (SSE) - является частью вариации переменной Y, которую невозможно описать с помощью регрессионной модели. Эта величина зависит от разностей между наблюдаемыми и предсказанными значениями.

Таким образом, оценками изменчивости в регрессионной модели являются:

Полная сумма квадратов (SST), равная сумме квадратов регрессии, плюс сумма квадратов ошибок

$$SST = SSR - SSE \quad (5)$$

Иначе, полная сумма квадратов (SST) может быть представлена как сумма квадратов разностей между наблюдаемыми значениями переменной Y и ее средним значением:

$$SST = \sum_{i=1}^{n} \left(Y_i - Y_{icp}\right)^2 \quad (6)$$

Сумма квадратов регрессии (SSR) равна сумме разностей между предсказанными значениями переменной Y и ее средним значением:

$$SSR = \sum_{i=1}^{n} \left(\acute{Y}_i - Y_{icp}\right)^2 \quad (7)$$

Сумма квадратов ошибок (SSE) равна сумме квадратов разностей между наблюдаемыми и предсказанными значениями:

$$SSE = \sum_{i=1}^{n} \left(Y_i - \acute{Y}_i\right)^2 \quad (8),$$

Суммы квадратов, вычисленные с помощью программы MS Excel по формулам (6 – 8), оказались равными: SSR=105.7476, SSE = 11.2967, SST = 116.9543

Нетрудно проверить по формуле (5), что полная сумма квадратов разностей SST и SSR равна 116.9543 (116.9543 = 105.7476 + 11.2967)

Следует обратить внимание на то, что в некоторых версиях программы MS Excel величина SSR представляется в так называемом научном формате. Этот формат применяется для представления очень маленьких или очень больших числовых величин [2,3]. Число, стоящее после буквы E, задает количество позиций, на которое следует перенести десятичную точку: влево – если это число отрицательное; вправо – если положительное. Например, запись 3,7431 + 02 означает, что десятичную точку следует перенести на две позиции вправо, т.е. число равно 374, 31. Запись 3,7431 – 02 означает, что десятичную точку нужно перенести на две позиции влево, т.е. число равно 0,037431.

Кроме того, следует учесть, что при записи чисел в научном формате количество значащих цифр, как правило, уменьшается, и числа могут округляться.

Заключение.

Выполненные решения многочисленных примеров и реальных задач с использованием MS Excel и PH Stat2 показали, что в диаграммах разброса и графиках остатков данные нередко отличаются друг от друга и иллюстрируют такую ситуацию, в которой эмпирическая модель значительно зависит от отдельного отклика [4]. Чтобы избежать ошибки и «подводные камни» при регрессионном анализе, необходимо:

- Помнить, что графики остатков являются необходимым инструментом регрессионного анализа и должны быть его неотъемлемой частью.
- Анализ возможной взаимосвязи между переменными X и Y всегда следует начинать с построения диаграммы разброса;
- Прежде чем интерпретировать резуль-таты регрессионного анализа, следует про-верять условия его применимости;
- Для определения соответствия эмпирической модели результатам наблюдения и обнаружения нарушены ли условия гомоскедастичности, целесообразно построить и исследовать график зависимости остатков от независимой переменной;
- Для проверки предположения о нормальном распределении ошибок следует использовать гистограммы, диаграммы «ствол и листья», блочные диаграммы и кривые нормального распределения (кривые Гаусса). При этом полезно оценить все основные моменты распределения кривой Гаусса – математическое ожидание, дисперсию, коэффициенты ассиметрии и эксцесс;
- Если условия применимости метода наименьших квадратов выполняются, необходимо проверить гипотезу о статистической значимости коэффициента регрессии и построить доверительные интервалы, содержащие математическое ожидание и предсказанное значение отклика;
- Если условия применимости метода наименьших квадратов не выполняются, следует использовать альтернативные методы, например, модели квадратичной или множественной регрессии;
- Не следует применять экстраполяцию, то есть необходимо избегать

предсказаний значения зависимой переменной за пределами изменения независимой переменной;

- Нужно иметь в виду, что корреляция между переменными не означает наличия причинно-следственной зависимости между ними, поскольку статистические зависимости не всегда являются причинно-следственными.

СПИСОК ЛИТЕРАТУРЫ

1. Левин, Дэвид М.., Стефан Дэвид, Кребиль, Тимоти С, Беренсон Марк Л. Статистика для менеджеров с использованием Microsoft Excel, 4-е изд. Пер. с англ.- Издательский дом "Вильямс" . 2004. - l312 с.: ил.

2. Джеффери Мур, Лоренс Р. Уэдерфорд, Ларри Р. Уэдерфорд. Экономическое моделирование в Microsoft Excel, 6-е изд. –М., 2004. - 102 с.

3. Норман Дрейпер, Гарри Смит. Прикладной регрессионный анализ. Множественная регрессия, - Applied Regression Analysis. - 3-е изд- М.: «Диалектика», 2007.- 912 с.: ил.

4. Крученецкий В.З., Серикулова Ж.К, Калабина А.А. О некоторых проблемах, связанных с использованием простой линейной регрессии и стандартных компьютерных средств при решении задач бизнес-процессов. /Материалы международной научно-практической конференции «Инновационное развитие пищевой, легкой промышленности и индустрии гостеприимства», 17-18 октября 2013, Алматы. –С. 362-328.