Структурированный по темам и разделам

лекционный теоретический материал

ЭКОНОМЕТРИКА

Авторы **Е. М. Кочкина**

Е. В. Радковская

Екатеринбург

2014

Большинство явлений и процессов в экономике находятся в постоянной взаимной и всеохватывающей объективной связи. Исследование зависимостей и взаимосвязей между существующими явлениями и процессами играет большую роль в экономике. Оно дает возможность глубже понять сложный механизм причинно-следственных отношений между явлениями.

Для исследования интенсивности, вида и формы зависимостей широко применяется корреляционно-регрессионный анализ, который является методическим инструментарием при решении задач прогнозирования, планирования и анализа хозяйственной деятельности.

Основная задача эконометрики состоит в построении моделей специфического типа (эконометрических моделей), описывающих взаимообусловленное развитие социально-экономических процессов на основе информации, отражающей распределений их уровней во времени и/или в пространстве однородных объектов. Эти модели используются в анализе и прогнозировании общих закономерностей и конкретных количественных характеристик рассматриваемых процессов, определении управляющих воздействий.

В результате, в самом широком толковании эконометрику можно рассматривать как объединений ряда дисциплин – экономической теории (включая микро- и макроэкономику, социальную сферу), математической и социально-экономической статистики, методов экономико-математического моделирования. Помимо вышеназванных дисциплин, одним из основных факторов развития эконометрики является развитие компьютерных технологий и специализированных пакетов прикладных программ.

Построение экономико-статистических моделей позволяет дать качественную и количественную характеристику зависимости между экономическими показателями. И хотя модель претендует лишь на упрощенное отражение действительности, она обеспечивает математический подход к исследованию сложившихся экономических взаимосвязей, к выяснению вопросов о том, существенна ли изучаемая зависимость, в какой форме она

проявляется и т.д. Экономико-статистическая модель служит не только средством анализа предшествующего экономического развития, но и становится важным инструментом плановых расчетов.

Задачи, решаемые с помощью эконометрики, классифицируются по трем признакам:

1) по конечным прикладным целям:

– прогнозирование социально-экономических показателей, которые характеризуют состояние и развитие изучаемой системы;

– моделирование возможных вариантов социально-экономического развития системы, в результате чего становится возможным определение параметров, которые оказывают наиболее сильное влияние на состояние системы в целом;

2) по уровню иерархии:

– задачи макроуровня (страна в целом);

– задачи мезоуровня (уровень отраслей, регионов);

– задачи микроуровня (уровень фирмы, семьи, предприятия);

3) по области решения проблем изучаемой экономической системы:

– задачи изучения рынка;

– задачи изучения инвестиционной, социальной, финансовой политики;

– задачи изучения ценообразования;

– задачи изучения распределительных отношений;

– задачи изучения спроса и потребления;

– задачи изучения отдельно выделенного комплекса проблем.

Решение перечисленных задач осуществляется с использованием математических моделей, построенных на основе эмпирических данных.

В эконометрике применяется два основных типа данных: пространственные и временные.

Пространственные данные – это совокупность экономической информации, характеризующей разные объекты и полученной за определенный период или в какой-то конкретный момент времени. Пространственные данные

являются выборкой из некоторой генеральной совокупности анализируемых данных (например, совокупность различной информации по какому-либо предприятию – численность работников, коэффициент сменности, процент текучести кадров, доля профильной продукции в объеме производства).

Временные данные – это совокупность экономической информации, характеризующей определенный объект, но за различные периоды времени. Отдельный временной ряд можно считать выборкой из бесконечного ряда значений показателей во времени (например, данные о динамике фондовых индексов).

В качестве переменных в эконометрических моделях могут рассматриваться разнообразные экономические показатели. В эконометрической модели используются:

– эндогенные (зависимые, результативные) переменные, также называемые объясняемыми переменными (y), их значения определяются внутри модели;

– экзогенные (факторные, независимые) переменные, также называемые объясняющими переменными (x), их значения задаются извне;

– лаговые (экзогенные или эндогенные) переменные, которые относятся к предыдущим моментам времени и находятся в одном уравнении с переменными, относящимися к текущему моменту времени.

Основная цель эконометрического моделирования – это характеристика значений одной или нескольких текущих эндогенных переменных в зависимости от значений объясняющих переменных.

Существует три основных класса эконометрических моделей.

1. Модели временных рядов представляют собой зависимость результативной переменной от переменной времени или переменных, относящихся к другим моментам времени.

Модели временных рядов, в которых результативная переменная зависит от времени:

– модель тренда (зависимость результативной переменной от трендовой компоненты);

– модель сезонности (зависимость результативной переменной от сезонной компоненты);

– модель тренда и сезонности.

Модели временных рядов, в которых результативная переменная зависит от переменных, датированных другими моментами времени:

– модели с распределенным лагом, объясняющие изменение результативной переменной в зависимости от предыдущих значений факторных переменных;

– модели авторегрессии, объясняющие изменение результативной переменной в зависимости от предыдущих значений результативных переменных;

– модели ожидания, объясняющие изменение результативной переменной в зависимости от будущих значений факторных или результативных переменных.

2. Регрессионные модели с одним уравнением, в которых результативная (зависимая) переменная может быть представлена в виде функции факторных (независимых) переменных:

$$y = f(x_1, x_2, \ldots, x_n, b_1, b_2, \ldots, b_n),$$

где $b_1, b_2, \ldots, b_n$ – параметры регрессионной модели.

По количеству факторных переменных регрессионные модели делятся на парные (с одной факторной переменной) и множественные регрессии (с двумя и более факторными переменными).

По виду функции $f(x_1, x_2, \ldots, x_n, b_1, b_2, \ldots, b_n)$ регрессионные модели делятся на линейные и нелинейные регрессионные модели.

3. Системы эконометрических уравнений.

Системы эконометрических уравнений предназначены для исследования тех экономических процессов, которые невозможно описать одним уравнением

регрессии. В этом случае строятся несколько эконометрических уравнений, которые в результате образуют систему.

В эконометрическом моделировании выделяют три вида систем уравнений:

1. Система независимых уравнений:

$$y_1 = a_{11}\, x_1 + a_{12}\, x_2 + \ldots + a_{1n}\, x_n + \acute{\varepsilon}_1$$

$$y_2 = a_{21}\, x_1 + a_{22}\, x_2 + \ldots + a_{2n}\, x_n + \acute{\varepsilon}_2$$

…

$$y_n = a_{n1}\, x_1 + a_{n1}\, x_2 + \ldots + a_{nn}\, x_n + \acute{\varepsilon}_n$$

Данная система уравнений характеризуется тем, что каждая эндогенная переменная *y* является функцией одних и тех же факторных переменных *x*.

2. Система рекурсивных уравнений:

$$y_1 = a_{11}\, x_1 + a_{12}\, x_2 + \ldots + a_{1m}\, x_m + \acute{\varepsilon}_1$$

$$y_2 = b_{21}\, y_1 + a_{21}\, x_1 + a_{22}\, x_2 + \ldots + a_{2m}\, x_m + \acute{\varepsilon}_2$$

$$y_3 = b_{31}\, y_1 + b_{32}\, y_2 + a_{n1}\, x_1 + a_{n2}\, x_2 + \ldots + a_{nm}\, x_m + \acute{\varepsilon}_n$$

…

$$y_n = b_{n1}\, y_1 + b_{n2}\, y_2 + \ldots + b_{nn\text{-}1}\, y_{n\text{-}1} + a_{n1}\, x_1 + a_{n2}\, x_2 + \ldots + a_{nm}\, x_m + \acute{\varepsilon}_n$$

Данная система уравнений характеризуется тем, что в каждом последующем уравнении эндогенная переменная y_i выступает в качестве экзогенной переменной.

3. Система взаимозависимых уравнений:

$$y_1 = b_{12}\, y_2 + b_{13}\, y_3 + \ldots + b_{1n}\, y_n + a_{11}\, x_1 + a_{12}\, x_2 + \ldots + a_{1m}\, x_m + \acute{\varepsilon}_1$$

$$y_2 = b_{21}\, y_1 + b_{23}\, y_3 + \ldots + b_{2n}\, y_n + a_{21}\, x_1 + a_{22}\, x_2 + \ldots + a_{2m}\, x_m + \acute{\varepsilon}_2$$

…

$$y_n = b_{n1}\, y_1 + b_{n2}\, y_2 + \ldots + b_{nn\text{-}1}\, y_{n\text{-}1} + a_{n1}\, x_1 + a_{n1}\, x_2 + \ldots + a_{nm}\, x_m + \acute{\varepsilon}_n$$

Данная система уравнений характеризуется тем, что эндогенные переменные в одних уравнениях входят в левую часть (т.е. являются результативными переменными), а в других уравнениях – в правую часть (т.е. являются факторными переменными).

В системе взаимозависимых уравнений значения результативных и факторных переменных формируются одновременно под влиянием внешних факторов. Поэтому данная система также называется системой одновременных, или совместных, уравнений.

В системах независимых и рекурсивных уравнений каждое уравнение может рассматриваться самостоятельно, и неизвестные коэффициенты таких уравнений можно определить классическим методом наименьших квадратов.

В системах взаимозависимых уравнений уравнения не могут рассматриваться как самостоятельная часть системы, поэтому применение традиционного метода наименьших квадратов для определения неизвестных коэффициентов таких уравнений невозможно.

Для решения эконометрической задачи необходимо последовательно выполнить несколько этапов экономико-математического моделирования.

Основу эконометрики составляет регрессионный анализ. Он применяется для того, чтобы при сложном взаимодействии посторонних влияний выяснить, какова была бы зависимость между результатом и фактором, если бы посторонние факторы не изменялись и своим изменением не искажали основную зависимость. Необходимо изучить выбранное явление во всех сложных взаимоотношениях с окружающими явлениями-факторами. При этом небольшое число наблюдений не дает возможности обнаружить закономерность связи.

В рамках регрессионного анализа нужно решить 4 основных задачи.

1. Определение числовых значений параметров модели;
2. Определение статистической достоверности параметров модели;
3. Расчет и анализ показателей качества построенной регрессионной модели;
4. Определение статистической достоверности построенной регрессионной модели.

Модели парной линейной регрессии

При изучении влияния одних признаков явления на другие из цепи признаков, характеризующих данное явление, выделяются два признака: факториальный и результативный. Необходимо установить, какой из них является факториальным и какой – результативным. В этом помогает, прежде всего, логический предметный анализ.

В простейшем случае исследуется связь между двумя показателями, из которых один рассматривается как независимый показатель-фактор[1], а второй – как зависимая переменная[2]. Наличие самой зависимости между этими показателями устанавливается, конечно, не математическим путем, а в результате качественного анализа, позволяющего вскрыть внутреннюю сущность изучаемого явления и порождающих его причин. Сам же регрессионный анализ предназначен для количественного измерения выявленной связи, хотя он нередко способствует и уточнению выводов самого качественного анализа.

Например, исследуется зависимость между стоимостью некоторого товара и объемом его продаж. Данные об изменении этих показателей собраны за определенный промежуток времени. Несомненно, анализируемые показатели связаны между собой. В качестве зависимой переменой (y) возьмем величину объема продаж, а в качестве независимой (x) – стоимость товара.

В первую очередь необходимо установить вид функции, связывающей показатели y и x, то есть найти такой вид уравнения регрессии, который наилучшим образом соответствует характеру изучаемой связи. Определение вида уравнения регрессии (вида связи) является важнейшей составной частью регрессионного анализа, поэтому его правильный подбор относится к наиболее ответственным этапам проводимого исследования.

Самый простой способ определения вида связи между показателями – визуальный. Нанесем на координатную плоскость все имеющиеся пары

[1] В дальнейшем этот показатель будет обозначаться через x.
[2] В дальнейшем этот показатель будет обозначаться через y.

наблюдений (рис.1): цена товара в период t (x_t) и объем продаж товара в период t (y_t). Полученный разброс точек на координатной плоскости называется *корреляционным полем*.

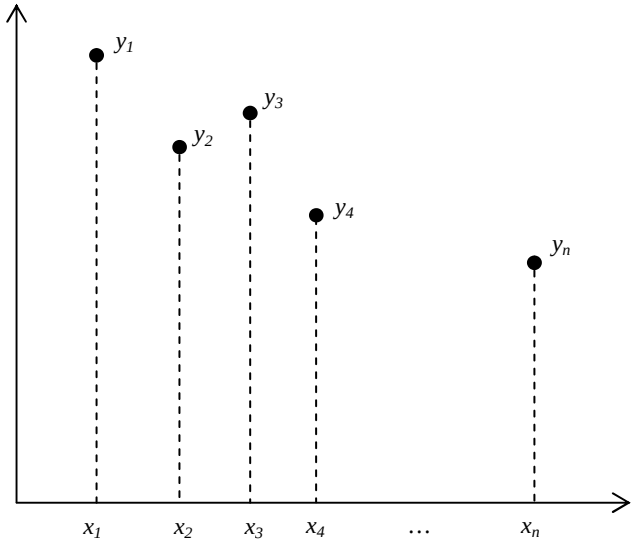

Рис. 1. Корреляционное поле

Если на корреляционном поле визуально не вырисовывается одна из нелинейных функций, то для моделирования связи можно использовать линейную зависимость. Большинство экономических процессов достаточно корректно описывается линейными (или кусочно-линейными) связями в основном диапазоне своих наблюдаемых значений.

Соответственно, простейшим уравнением, которое может характеризовать зависимость между двумя переменными, является линейное уравнение. Предположим, что связь между анализируемыми показателями является именно линейной, то есть описывается уравнением прямой вида:

$$y_t = \alpha + \beta x_t + \varepsilon_t \qquad (1),$$

где x_t и y_t – соответственно независимая и зависимая переменные, α и β – коэффициенты регрессии, а ε_t – случайная компонента, характеризующая ошибки – возможные отклонения между реальными и расчетными значениями зависимой переменной y_t.

Сразу же отметим, что не следует ожидать получения точного соотношения между какими-либо двумя (или – в общем случае – более) экономическими показателями, за исключением тех случаев, когда оно

существует по определению. В статистическом анализе факт неточности соотношения признается путем явного включения в него случайного фактора, описываемого случайной составляющей ε_t (остаточным членом).

При этом полагается, что

x_t – неслучайная детерминированная величина, ее называют объясняющей (независимой) переменной, или регрессором (фактором);

y_t, ε_t – случайные величины;

y_t – объясняемая зависимая переменная (результирующий показатель);

ε_t – величина, характеризующая влияние на результирующий показатель неучтенных в модели факторов.

Очевидно, что чем меньше значения ε_t, тем точнее решается первая задача регрессионного анализа, которая состоит в получении оценок α и β.

Решение первой задачи регрессионного анализа

Предположим, что у нас имеется n наблюдений для x_t и y_t, и перед нами стоит задача – определить значения α и β в уравнении $y_t = \alpha + \beta x_t + \varepsilon_t$. Зависимость $\alpha + \beta x_t$ – это уравнение прямой линии на плоскости, заданное в общем виде.

С самого начала необходимо признать, что мы никогда не сможем рассчитать истинные значения α и β. Можно получить только их оценки (a и b), которые могут быть хорошими или плохими. (Иногда оценки бывают абсолютно точными, но это возможно лишь в результате случайного совпадения, и даже в этом случае нет способа узнать, что оценки абсолютно точны).

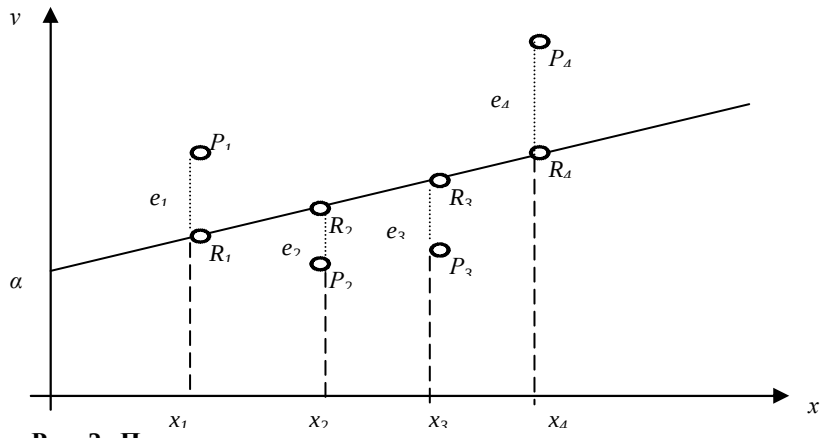

Рис. 2. Построение линии регрессии на корреляционном поле

Отрезок, отсекаемый прямой на оси *y*, представляет собой оценку α и обозначен *a*, а угловой коэффициент прямой представляет собой оценку β и обозначен *b*.

Вообще говоря, через корреляционное поле можно провести бесконечное множество прямых линий. Среди этого множества нас интересует та, точки которой наилучшим образом согласуются с реальными данными. Чтобы найти параметры интересующей нас линии (линии регрессии) используется специальный математический метод, называемый *методом наименьших квадратов* (подробно с реализацией метода наименьших квадратов можно ознакомиться по предлагаемой литературе в рабочей программе).

Метод наименьших квадратов наиболее широко применяется при решении задач регрессионного анализа. С его помощью находятся такие числовые значения коэффициентов *a* и *b*, для которых сумма квадратов отклонений между реальными и модельными значениями зависимой переменной *y* была минимальной.

Экономико-математическая интерпретация построенной регрессионной модели

После записи уравнения регрессии необходимо выполнить экономико-математическую интерпретацию полученной модели. Модель парной линейной регрессии в общем случае имеет вид:

$$\hat{y} = a + bx.$$

Формально параметр a дает прогнозируемое значение y при нулевом значении x. Однако в экономических задачах показатель x редко принимает нулевое значение и буквальная интерпретация может привести к неверным результатам.

Поэтому в процессе интерпретации эконометрической модели основное внимание следует уделять не величине, а знаку параметра a, который здесь определяет относительную скорость изменения показателей, включенных в модель. Если $a > 0$, то относительное изменение x происходит быстрее, чем изменение y. Если $a < 0$, то относительное изменение y происходит быстрее, чем изменение x.

Проиллюстрируем смысловую интерпретацию коэффициента регрессии b. Предположим, что величина показателя x увеличилась на 1 единицу, тогда:

$$\hat{y} = a + b\,(x + 1) = a + bx + b.$$

Т.о., видно, что увеличение x на 1 единицу приводит к изменению зависимой переменной y на величину b. Важную роль в интерпретации коэффициента b играет его знак. Если $b > 0$, с ростом x растет y, и связь между показателями является прямой. Если $b < 0$, с ростом x величина y падает, и связь между показателями является обратной.

Решение второй задачи регрессионного анализа

Вторая задача регрессионного анализа заключается в проверке статистической достоверности параметров построенной регрессионной модели. Математически параметры a и b можно рассчитать для любого набора статистической информации, однако необходимо проверить, можно ли доверять найденным значениям.

Величины a, b и e_t зависят от величины ошибки ε_t. Остатки e_t оценивают величину ошибки ε_t, но, в отличие от нее, являются наблюдаемыми.

Величина ошибки рассчитывается на основе дисперсии. Точный расчет дисперсии ошибок, как правило, невозможен, так как нами исследуется ограниченный набор данных. Однако, если взятая выборка является

представительной, то оценить дисперсию ошибок ε_t можно по величине остатков e_t, которые представляют собой разности между реальными и модельными значениями зависимой переменной.

Если стандартная ошибка коэффициента регрессии не превышает половины модуля самого коэффициента, то найденный коэффициент можно признать значимым (достоверным).

Применятся и другой способ оценки значимости найденных коэффициентов регрессии, основанный на использовании статистических гипотез.

Проверка статистической гипотезы о достоверности параметра b складывается из следующих этапов:

1) выдвигается нулевая гипотеза $H_0(b)$: $b = 0$, согласно которой коэффициент b будет равен нулю при неограниченном увеличении объема статистической информации, а при анализе имеющегося ограниченного набора статистических данных получился не равным нулю;

2) необходимо определить, существенно ли найденное значение параметра b отличается от нуля. В качестве базиса для проверки используются имеющиеся статистические данные. Для этого необходимо ввести такую переменную, по значению которой можно было бы судить о справедливости нулевой гипотезы. Такой переменной является статистика Стьюдента, обозначаемая t:

$$t = \frac{b}{\sigma_b}.$$

3) по таблице распределения Стьюдента определяется значение t-статистики, которое является критическим значением для оцениваемого коэффициента регрессии. Если значение анализируемого коэффициента регрессии по модулю больше значения t-статистики для него, то нулевая гипотеза отвергается. В противном случае нулевую гипотезу отвергнуть нельзя. Это не означает, что мы ее принимаем, мы только не можем ее отвергнуть и, следовательно, нужны дополнительные исследования.

В большинстве случаев определяется не только величина статистики Стьюдента, но и рассчитанная на ее основе вероятность выполнения нулевой гипотезы. Нулевая гипотеза отвергается, если вероятность ее выполнения меньше 5%. Если данная вероятность больше или равна 5%, нуль-гипотезу отвергнуть нельзя и, следовательно, между x_t и y_t нет линейной связи, а иногда делается не вполне строгий вывод о том, что изменение y_t не зависит от изменения x_t.

Полностью аналогично проверяется выполнение нулевой гипотезы для параметра a. Если нулевую гипотезу для параметра a нельзя отвергнуть, то зависимость между x_t и y_t превращается в простую пропорциональную зависимость.

Однако точечной оценки для параметров α и β недостаточно. Важно определить, в какой интервал в большинстве случаев (в 95% случаев) будут попадать истинные значения параметров α и β при изменении набора анализируемых данных. Зная табличное значение статистики Стьюдента ($t_{табл}$), можно определить границы искомых интервалов.

Для параметра α: $[a - t_{табл} \cdot \sigma_a \, , \, a + t_{табл} \cdot \sigma_a]$.

Для параметра β: $[b - t_{табл} \cdot \sigma_b \, , \, b + t_{табл} \cdot \sigma_b]$.

Записанные интервалы называются доверительными интервалами с 95%-м уровнем доверия.

Решение третьей задачи регрессионного анализа

Решение третей задачи предполагает расчет и анализ показателей качества построенной регрессионной модели.

Мы предположили, что показатели x_t и y_t связаны между собой, более того, предположили, что эта связь линейная, то есть определили форму связи между изучаемыми показателями, нашли параметры a и b и оценили их статистическую значимость.

Теперь необходимо установить, насколько эта связь является тесной. Регрессия характеризует зависимость изменения среднего значения фактора y_t от изменения фактора x_t. Для любого из имеющихся наблюдений справедливо:

$$y_t = a + bx_t + e_t = \hat{y}_t + e_t \quad (3).$$

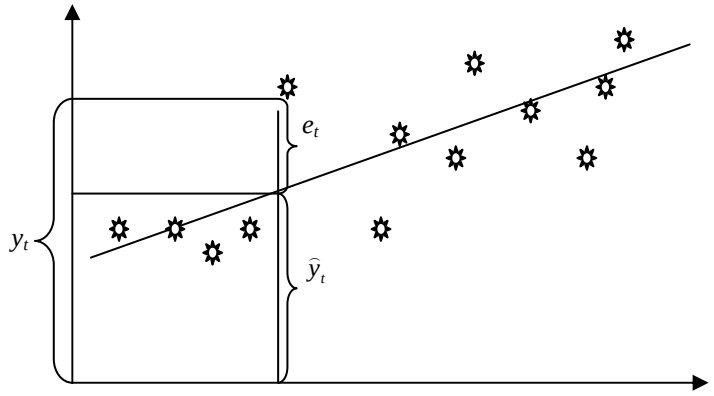

Рис. 3. Графическое представление составляющих реального значения зависимой переменной

Из уравнения (3) видно, что реальное значение зависимой переменной равно сумме модельного значения зависимой переменной и остатка, что показано на рис. 3.

Вычтем из обеих частей равенства (3) значение $\bar{y}$. В результате получим:

$$y_t - \bar{y} = \hat{y}_t - \bar{y} + e_t \quad (4).$$

Возведя в квадрат обе части равенства (4) и просуммировав полученные результаты по *t* от 1 до *n*, полагая, что $\sum_{t=1}^{n} e_t = 0$ получим:

$$\sum_{t=1}^{n}(y_t - \bar{y})^2 = \sum_{t=1}^{n}(\hat{y}_t - \bar{y})^2 + \sum_{t=1}^{n} e_t^2 \quad (5).$$

Введем следующие обозначения:

$\sum_{t=1}^{n}(y_t - \bar{y})^2 = TSS$ – сумма квадратов отклонений реальных значений зависимой переменной от ее среднего значения или общая дисперсия.

$\sum_{t=1}^{n}(\hat{y}_t - \bar{y})^2 = RSS$ – сумма квадратов отклонений модельных значений зависимой переменной от ее среднего значения или объясненная дисперсия.

$\sum_{t=1}^{n} e_t^2 = ESS$ – сумма квадратов отклонений реальных значений зависимой переменной от ее модельных значений (сумма квадратов остатков) или остаточная (необъясненная) дисперсия.

Формулу (5) можно переписать в виде:

$$TSS = RSS + ESS \quad (6).$$

В равенстве (6) разделим все элементы на величину *TSS*:

$$\frac{TSS}{TSS} = 1 = \frac{RSS}{TSS} + \frac{ESS}{TSS}.$$

Величина $\frac{ESS}{TSS}$ убывает при более правильном выборе линии регрессии, и наоборот. Так как эта величина всегда сверху ограничена единицей, то вместо нее в качестве показателя согласия используют величину:

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

$$0 \le R^2 \le 1.$$

Величина R^2 называется *коэффициентом детерминации* и показывает, какая доля вариации зависимой переменной может быть объяснена уравнением регрессии.

Если предположить, что вся вариация в y_t полностью определяется случайными возмущениями и не связана с изменением x_t, тогда *RSS*=0, а *ESS*=*TSS*, то есть R^2 =0.

Коэффициент детерминации показывает качество «подгонки» регрессионной модели к значениям y_t.

В качестве меры степени тесноты линейной связи переменных (в нашем случае x_t и y_t) часто используется *коэффициент корреляции R*.

Функционально коэффициент корреляции связан с коэффициентом детерминации и рассчитывается как корень квадратный из последнего. В отличие от коэффициента детерминации, изменяющегося от 0 до 1, в случае парной регрессии, коэффициент корреляции принимает значения на отрезке [–1,+1]. При этом:

1) если значение коэффициента корреляции меньше нуля, то связь между изучаемыми показателями x_t и y_t является обратной, т.е., с увеличением x_t значение y_t уменьшается, и наоборот;

2) если значение коэффициента корреляции больше нуля, то связь между изучаемыми показателями x_t и y_t является прямой, т.е., с увеличением x_t значение y_t увеличивается;

3) если значение коэффициента корреляции равно нулю, то линейная связь между изучаемыми показателями x_t и y_t отсутствует;

4) если значение коэффициента корреляции равно +1 или –1, то линейная связь между изучаемыми показателями x_t и y_t является строго функциональной, т.е., изменение факторного признака x_t полностью определяет изменение результативного признака y_t.

Близкий к 0 коэффициент корреляции говорит об отсутствии линейной зависимости между переменными. В этом случае может присутствовать нелинейная связь переменных, либо зависимость вообще отсутствует.

Существуют многочисленные исследования зависимости величины коэффициента корреляции от количества наблюдений. Не углубляясь в суть этих исследований, отметим, что связь между x_t и y_t считается тесной, когда коэффициент корреляции по модулю больше или равен 0,7. Условная градация величины коэффициента корреляции приведена в таблице 1.

Таблица 1.Характеристика тесноты связи в зависимости от величины коэффициента корреляции

Величина коэффициента корреляции	*Теснота связи между показателями в модели*
$\lvert R \rvert = 0$	связь отсутствует
$0 < \lvert R \rvert < 0{,}3$	связь слабая
$0{,}3 \leq \lvert R \rvert < 0{,}7$	связь средняя
$0{,}7 \leq \lvert R \rvert < 1$	связь тесная
$\lvert R \rvert = 1$	связь функциональная

Если на уровне теоретического исследования связь между показателями установлена, а расчетное значение коэффициента корреляции $R < 0{,}7$, то целесообразно использовать одну или несколько из нижеперечисленных процедур:

1) добавить в регрессионную модель новые регрессоры, поскольку результирующий показатель y_t может реально зависеть не только от x_t, но и от других факторов;

2) перейти к нелинейной регрессионной модели, т.к. некоторые экономические процессы не могут быть адекватно описаны линейной моделью;

3) удалить из анализируемой статистики статистические выбросы.

Статистический выброс – это аномальное наблюдение, для которого реальное значение результирующего показателя y_t резко отклоняется от линии регрессии. Чаще всего такое отклонение обусловлено единовременным воздействием случайных факторов, не характерных для изучаемого процесса. Присутствие в анализируемой статистике наблюдений-выбросов искажает полученные результаты. После выявления наблюдений-выбросов их следует удалить из анализируемых данных. При этом следует помнить, что количество удаляемых наблюдений не должно превышать 1/8 общего объема данных.

При использовании регрессионного анализа для динамических рядов не следует удалять последнее наблюдение. Если последнее наблюдение является статистическим выбросом, то следует дождаться поступления новых данных и убедиться, что они будут располагаться достаточно близко к линии регрессии. Если последующие наблюдения не приближаются к линии регрессии, то можно сделать вывод о том, что изучаемый процесс вследствие каких-либо причин стал развиваться по иному закону и построенную регрессионную модель нельзя использовать для его дальнейшего исследования.

Решение четвертой задачи регрессионного анализа

Четвертая задача регрессионного анализа состоит в определении статистической достоверности построенной модели.

Это, в первую очередь, подразумевает оценку статистической достоверности коэффициента детерминации. Здесь опять встает проблема выбора переменной, с помощью которой можно было бы судить о справедливости предположения, что связь между показателями x_t и y_t

отсутствует и что $R^2=0$ при неограниченном увеличении объема статистических данных.

Величина, с помощью которой проверяется нулевая гипотеза для коэффициента детерминации, называется статистикой Фишера.

Величина F подчиняется распределению Фишера (F-распределение), зная которое можно рассчитанную статистику Фишера сравнить с табличным значением. Случайное превышение табличного значения маловероятно.

Если $F_{табличное} < F_{фактическое}$, то нулевая гипотеза для коэффициента детерминации отвергается, т.е., вариация y_t обусловлена не только случайными возмущениями, но и вариацией x_t. Если $F_{табличное} > F_{фактическое}$, то нулевую гипотезу для коэффициента детерминации отвергнуть нельзя. Это не означает, что x_t не влияет на y_t, просто на анализируемых статистических данных это влияние установить не удалось.

По распределению Фишера можно определить вероятность выполнения нулевой гипотезы для коэффициента детерминации. Логика проверки выстраивается следующим образом:

1) выдвигается нуль-гипотеза, согласно которой R^2 в действительности равен нулю, а его расчетное значение отлично от нуля из-за ограниченности имеющегося набора статистических данных;

2) определяется статистика Фишера, имеющая F-распределение;

3) по распределению статистики Фишера рассчитывается вероятность выполнения гипотезы H_0:

а) если вероятность больше или равна 5%, то нулевую гипотезу отвергнуть нельзя, установленная линейная связь между x_t и y_t не является статистически достоверной, необходимо увеличить количество наблюдений;

б) если вероятность меньше 5%, то нулевая гипотеза отвергается на 95%-м уровне значимости, найденному значению коэффициента детерминации можно доверять, а размер используемой выборки признать достаточным.

Пример построения и анализа модели парной линейной регрессии с использованием пакета Microsoft Excel

Решение эконометрических задач с помощью МНК реализовано во многих пакетах прикладных программ. Достаточно удобный интерфейс для этого предусмотрен в Microsoft Excel, но который мы и будем ориентироваться при описании решения задач регрессионного анализа.

Рассмотрим практические подходы к построению и анализу эконометрических моделей на примере конкретной задачи.

Менеджер новой чебуречной не уверен в правильности выбранной цены на чебуреки, поэтому на протяжении определенного времени он варьирует цену и отслеживает количество проданных чебуреков. Статистические данные приведены в таблице.

Ставятся следующие задачи:

1) построить эконометрическую модель зависимости количества проданных чебуреков от цены;

2) исследовать качественные характеристики построенной эконометрической модели;

3) на основе модели определить оптимальную в смысле максимума выручки цену чебурека.

На рис. 4 представлены собранные менеджером статистические данные, занесенные в электронную таблицу Microsoft Excel.

	A	B	C
6			
7	Номер недели	Цена	Объем
8	1	10,69	1196
9	2	9,30	1307
10	3	8,74	1433
11	4	13,78	751
12	5	13,89	481
13	6	11,68	1009
14	7	12,32	774
15	8	10,39	1127
16	9	11,44	1002
17	10	10,22	1254
18	11	13,74	578
19	12	12,35	665
20	13	10,77	1157
21	14	12,76	696
22	15	8,32	1638
23	16	8,31	1571
24	17	11,76	966
25			

Рис. 4. Таблица исходных данных для построения модели

Для получения численного решения задачи в Microsoft Excel следует воспользоваться программой анализа данных стандартного Пакета анализа. В Excel 2007 для выполнения регрессионного анализа необходимо последовательно выбрать следующие пункты меню: *Данные → Анализ данных → Регрессия*.

При отсутствии опции *Анализ данных* в меню *Данные* Пакет анализа следует подгрузить с помощью опции *Надстройки*.

В результате на экран вызывается окно диалога, которое необходимо заполнить:

Входной интервал Y: выделяются все значения зависимой переменной вместе с названием (в нашем случае, объем), т.е. выделяются ячейки, в которых содержатся числовые значения объема (в нашем случае, C7:C24);

Входной интервал X: выделяются все значения независимой переменной вместе с названием (в нашем случае, цена) , т.е. выделяются ячейки, в которых содержатся числовые значения цены (в нашем случае, B7:B24);

В позиции *Метки* ставится флажок, т.к. во входные интервалы включены не только числовые значения, но и имена переменных;

Помимо диапазона входных данных, задается информация о параметрах вывода. Результаты регрессионного анализа могут быть выведены на текущий рабочий лист, на отдельный рабочий лист (установкой флажка возле опции

Новый рабочий лист), в новый файл (установкой флажка возле опции *Новая рабочая книга*) или на текущий рабочий лист (установкой флажка возле опции *Выходной интервал*). В последнем случае результаты решения будут выведены на тот же лист, где находятся исходные данные задачи, начиная с той позиции, которая будет указана пользователем в поле *Выходной интервал*.

В нашем случае в параметрах вывода в позиции *Выходной интервал* указывается адрес ячейки, являющейся левой верхней ячейкой диапазона вывода результатов (ячейка А30);

В позициях *Остатки* и *Стандартизованные остатки* ставятся флажки, поскольку эти результаты необходимы для полноценного анализа.

Вид окна диалога в нашем случае представлен на рис. 5.

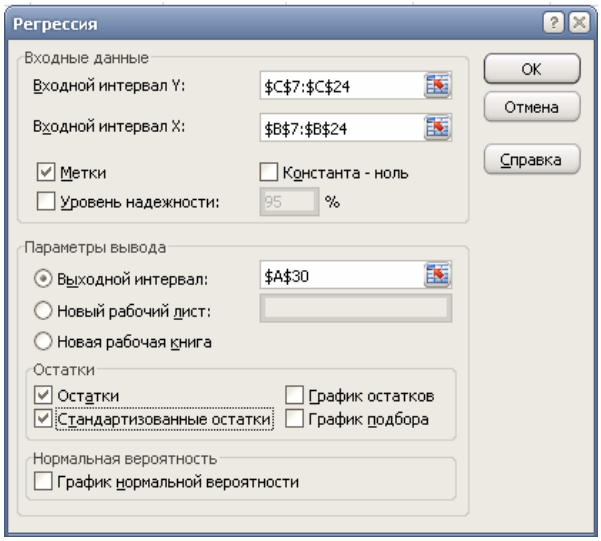

Рис. 5. Окно диалога «Регрессия»

После заполнения всех необходимых полей диалога и нажатия кнопки «ОК» на экран будет выведено решение задачи (рис.6).

	A	B	C	D	E	F	G
28	ВЫВОД ИТОГОВ						
30	*Регрессионная статистика*						
31	Множественный R	0,973					
32	R-квадрат	0,947					
33	Нормированный R-квадрат	0,943					
34	Стандартная ошибка	82					
35	Наблюдения	17					
36							
37	Дисперсионный анализ						
38		*df*	*SS*	*MS*	*F*	*Значимость F*	
39	Регрессия	1	1802215	1802215	267	0%	
40	Остаток	15	101191	6746			
41	Итого	16	1903406				
42							
43		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
44	Y-пересечение	3076	126	24	0	2807	3346
45	Цена	-182	11	-16	0	-206	-158
46							
47	ВЫВОД ОСТАТКА						
49	*Наблюдение*	*Предсказанное Объем*	*Остатки*	*Стандартные остатки*			
50	1	1129	66,88	0,84			
51	2	1382	-75,29	-0,95			
52	3	1484	-51,29	-0,64			
53	4	566	184,69	2,32			
54	5	546	-65,28	-0,82			
55	6	949	60,20	0,76			
56	7	832	-58,24	-0,73			
57	8	1184	-56,76	-0,71			
58	9	993	9,48	0,12			
59	10	1215	39,27	0,49			
60	11	574	4,40	0,06			
61	12	827	-161,77	-2,03			
62	13	1115	42,45	0,53			
63	14	752	-56,10	-0,71			
64	15	1561	77,21	0,97			
65	16	1563	8,39	0,11			
66	17	934	31,77	0,40			

Рис. 6. Результаты решения задачи в пакете Excel

Результаты проведенного регрессионного анализа выводятся в четырех таблицах под общим названием *Вывод итогов*.

Для полноценного анализа полученных результатов необходимо решить четыре выше описанные задачи регрессионного анализа.

Задача 1. Определение числовых значений коэффициентов модели.

Регрессионная модель в общем случае имеет вид:

$y = a + b\,x$.

В нашей задаче объем продаж выполняет роль зависимой переменной y, а цена одного чебурека является факторной переменной x. Для записи регрессионной модели в частном (а не общем) виде необходимо определить числовые значения параметров a и b.

Числовые значения параметров a и b выведены в столбце *Коэффициенты* третьей таблицы *Вывода итогов*.

Здесь *Y-пересечение* является константой уравнения регрессии (параметр a). В нашем случае константа a равна 3076.

Формально параметр *a* определяет величину *y* при нулевом значении *x*. Т.е. условно можно сказать, что при нулевой цене чебурека объем их продаж составит 3076 штук.

Однако поскольку в реальности цена не может иметь нулевого значения, то буквальная трактовка параметра *a* не имеет экономического смысла.

Проанализируем знак параметра *a*. В случае, когда $a > 0$, относительное изменение результата происходит медленнее, чем изменение фактора. В нашем случае *a* равно 3076, поэтому относительное изменение объема (*y*) происходит медленнее, чем изменение цены (*x*).

Коэффициент регрессии *b* показывает величину изменения результата при увеличении фактора на единицу. В случае, когда $b < 0$, связь между показателями является обратной, т.е.

с ростом *x* снижается *y*. В нашем случае *b* равно –182, следовательно, с ростом цены на 1 р. объем продаж снижается в среднем на 182 чебурека.

Выполнив анализ коэффициентов, можно записать полученную регрессионную модель:

Объем = 3075 – 182 Цена.

Коэффициент регрессии *b* дает возможность оценить, как в среднем меняется результирующий показатель при изменении фактора. Но поскольку, как уже отмечалось, в эконометрических исследованиях редко имеется возможность использовать генеральную совокупность данных, необходимо проанализировать границы изменения найденных коэффициентов.

Для этого рассмотрим доверительные интервалы коэффициентов, выведенные в столбцах *Нижние 95%* и *Верхние 95%*. Доверительный интервал показывает интервал изменения соответствующего параметра регрессии в 95% случаев при тех или иных изменениях исходных данных. В нашем примере величина константы при изменении исходных данных почти наверняка (с вероятностью 95%) будет лежать в интервале от 2807 до 3346, а величина коэффициента перед переменной *Цена* – в интервале от –206 до –158.

Т.о., в среднем при росте цены на 1 р., объем продаж снижается на 182 чебурека.

В лучшем случае объем продаж снизится на 158 чебуреков при увеличении цены на 1 р.

В худшем случае объем продаж снизится на 206 чебуреков при увеличении цены на 1 р.

В выводимых Excel результатах регрессии столбцы *Нижние 95%* и *Верхние 95%* повторяются дважды. Это связано с тем, что пользователю предоставляется возможность, помимо стандартного 95%-го указать интересующий его уровень значимости. Если есть необходимость помимо 95%-го доверительного интервала получить интервал с другим уровнем надежности результатов следует при заполнении окна диалога *Регрессия* нажатием левой кнопки мыши поставить флажок возле опции *Уровень надежности* и в соответствующем поле указать его значение.

Задача 2. Анализ статистической значимости коэффициентов регрессионной модели.

В рамках решения второй задачи для исследования значимости параметров также проверяется вероятность выполнения нулевой гипотезы для найденных коэффициентов a и b. Вероятность выполнения нулевой гипотезы проверяется с использованием статистики Стьюдента, числовые значения которой приводятся в столбце *t-статистика*.

Для нашего примера статистика Стьюдента для параметра a составляет 24, для параметра b равна –16.

На основании этих значений рассчитываются вероятности выполнения нулевых гипотез для обоих параметров, которые выводятся в столбце *P-Значение*.

В нашем случае вероятность выполнения нулевой гипотезы для коэффициента a (т.е. вероятность того, что $a = 0$) равна нулю (меньше порогового значения в 5%). Т.о., можно считать параметр a отличным от нуля и статистически достоверным. Вероятность выполнения нулевой гипотезы для

коэффициента b (т.е. вероятность того, что $b = 0$) также равна нулю. Т.о., параметр b тоже можно считать отличным от нуля и статистически достоверным.

Обобщая вышесказанное, подчеркнем, что *P-Значение* определяет:

–вероятность выполнения нулевой гипотезы для соответствующего коэффициента регрессии;

–т.е. вероятность незначимости (недостоверности) соответствующего коэффициента регрессии;

–т.е. вероятность того, что фактор x не оказывает линейного влияния на результативный показатель y.

Задача 3. Расчет и анализ показателей качества построенной регрессионной модели.

Расчет показателей качества модели проводится на основе дисперсионного анализа.

В таблице *Дисперсионный анализ* (вторая таблица *Вывода итогов*) в столбце *SS* указаны значения дисперсий: объясняемой регрессионной моделью (*RSS*), остаточной (*ESS*) и общей (*TSS*).

TSS – это сумма квадратов отклонений реальных значений y от среднего значения y. Величина *TSS* выводится в строке *Итого*. В нашем случае величина *TSS* равна 1 903 406.

RSS – это сумма квадратов отклонений модельных значений y от среднего значения y. Величина *RSS* выводится в строке *Регрессия*. В нашем случае величина *RSS* равна 1 802 215.

ESS – это сумма квадратов отклонений реальных значений y от модельных значений y. Величина *ESS* выводится в строке *Остаток*. В нашем случае величина *ESS* равна 101 191.

Как описывалось выше, зная значение дисперсий, можно рассчитать один из показателей качества регрессионной модели – коэффициент детерминации. Числовое значение коэффициента детерминации выводится в первой таблице *Вывода итогов* в строке *R-квадрат*.

R-квадрат – коэффициент детерминации, рассчитывается как отношение объясненной дисперсии (*RSS*) к общей дисперсии (*TSS*). В нашем случае:

$$R^2 = \frac{RSS}{TSS} = \frac{1802215}{1903406} = 0{,}947.$$

R-квадрат определяет:

– долю дисперсии, объясненную регрессионной моделью;

– долю разброса данных, объясненного регрессионной моделью;

– долю наблюдений, попавших под описание регрессионной модели.

Для решаемой задачи доля объясненной дисперсии составляет 94,7%, т.е. под описание регрессионной модели попадает 94,7% наблюдений.

Вычислив квадратный корень из коэффициента детерминации, получаем коэффициент корреляции, который измеряет тесноту связи в регрессионной модели и выводится в строке *Множественный R* первой таблицы *Вывода итогов*.

Коэффициент корреляции является важнейшим показателем для оценки качества регрессионной модели. Этот показатель определяет, насколько тесно связаны между собой зависимая и факторная переменные в построенной модели.

В нашем случае значение коэффициента корреляции близко к единице (*Множественный R* = 0,973), что свидетельствует о наличии достаточно тесной связи между исследуемыми экономическими показателями и подтверждает влияние цены чебурека на изменение объема продаж.

В некоторых случаях низкое значение коэффициента корреляции, как уже отмечалось выше, может быть связано с наличием в изучаемой выборке аномальных наблюдений – статистических выбросов, которые искажают как величины коэффициентов регрессии, определяющих меру влияния фактора на результат, так и характеристику тесноты связи. Если величина коэффициента корреляции в решаемой задаче меньше 0,7, то, в первую очередь, следует проверить наличие статистических выбросов. Если объем статистической выборки позволяет ее сократить, то даже при приемлемом значении

коэффициента корреляции рекомендуется исключать из набора исходных данных статистические выбросы.

Напомним, что статистический выброс – это наблюдение, резко отклонившееся от линии регрессии вверх или вниз. Если наблюдение является статистическим выбросом, его стандартный остаток по абсолютной величине больше или равен 2. Величины стандартных остатков выводятся в столбце *Стандартные остатки* четвертой таблицы *Вывода итогов*.

В нашем случае 4-е и 12-е наблюдения являются статистическими выбросами. С определенной долей условности можно считать, что для остальных наблюдений реальный и модельный объем продаж приблизительно совпадают.

4-е наблюдение является выбросом вверх (стандартный остаток = 2,32). Это не означает, что объем продаж в 4-м периоде был слишком большим. Появление этого выброса связано с тем, что при установленной в 4-м периоде цене объем продаж был значительно выше, чем можно было бы ожидать согласно построенной модели.

12-е наблюдение является выбросом вниз (стандартный остаток = –2,03). Это также не означает, что объем продаж в 12-м периоде был мал. Появление этого выброса связано с тем, что при установленной в 12-м периоде цене объем продаж был существенно ниже, чем можно было бы ожидать согласно построенной модели.

В нашем случае исходный объем выборки составлял 17 наблюдений, что позволяет удалить два обнаруженных выброса (17 / 8 > 2). Процедура удаления статистических выбросов заключается в удалении из исходных данных тех строк, которые соответствуют наблюдениям-выбросам (в нашем случае это 11 и 19 строки).

Величины остатков выводятся в столбце *Остатки* четвертой таблицы *Вывода итогов*.

По величине остатков можно сравнить реальные и модельные значения зависимой переменной. Если остаток для какого-либо наблюдения больше

нуля, то реальное значение в этом наблюдении больше модельного, и наоборот: при отрицательном остатке модельное значение больше.

Например, для рассматриваемой задачи в 6-м наблюдении реальное значение объема продаж меньше модельного, т.к. остаток для этого наблюдения равен –47,82. В 10-м наблюдении реальное значение объема продаж больше модельного на 24,07, т.к. остаток для этого наблюдения равен 24,07.

В четвертой таблице *Вывода итогов*, помимо остатков и стандартных остатков, выводятся модельные значения зависимой переменной для каждого наблюдения. Для расчета модельных значений в построенной регрессионной модели фактор *x* должен последовательно принять все реальные значения из изучаемой выборки. Модельные значения зависимой переменной *y* выводятся в столбце *Предсказанное y*.

Для решаемой задачи столбец имеет название *Предсказанное Объем* и содержит модельные значения объема продаж для всех 17-ти наблюдений. Например, в 15-м наблюдении модельный объем продаж равен 1561. Это значение можно получить подстановкой в регрессионную модель значения цены для 15-го периода:

Объем = 3076 – 182· Цена = 3076 – 182· 8,32 = 1561.

Задача 4. Определение статистической достоверности построенной регрессионной модели.

Статистическая достоверность регрессионной модели проверяется с помощью нулевой гипотезы для коэффициента детерминации. Используем величину статистики Фишера, которая выведена во второй таблице *Вывода итогов* в столбце *F*. В нашем случае величина *F* равна 540.

С помощью статистики Фишера определяется вероятность выполнения нуль-гипотезы для коэффициента детерминации, которая выводится в столбце *Значимость F*.

В нашей задаче *Значимость F* равна нулю, следовательно, нулевая гипотеза отвергается на 95%-м уровне значимости, а коэффициент детерминации признается статистически достоверным.

Значимость F определяет:

– вероятность выполнения нулевой гипотезы для коэффициента детерминации R^2;

– т.е. вероятность того, что наблюдений для проведения регрессии недостаточно.

После решения всех четырех задач регрессионного анализа делается общий вывод о качестве построенной модели. Особое внимание при оценке качества необходимо уделить следующим аспектам анализа:

1. связь между изучаемыми показателями должна быть тесной, т.е. коэффициент корреляции (*Множественный R*) должен быть больше или равен 0,7;

2. коэффициенты модели, определяющие меру влияния факторов на результат, должны быть достоверными, т.е. все *P-Значения* должны быть меньше 5%;

3. регрессионная модель в целом должна быть достоверна (количество наблюдений должно быть достаточным), т.е. величина *Значимость F* должна быть меньше 5%;

4. результаты регрессионного анализа не должны содержать статистических выбросов, которые могут быть удалены.

После удаления статистических выбросов построенную регрессионную модель можно признать качественной. Однако перед нами еще стояла задача нахождения оптимальной цены чебурека, при которой достигается максимум выручки.

Для измерения совместного влияния ряда показателей-факторов на величину анализируемого показателя строятся модели множественной регрессии.

Множественная регрессия широко используется в решении проблем спроса, доходности акций, при изучении функций издержек производства, в макроэкономических расчетах и в других эконометрических задачах. Основная цель множественной регрессии заключается в построении модели с большим числом факторов. При этом необходимо определить влияние каждого фактора в отдельности на результирующий показатель, а также их совокупное воздействие на него.

Для получения надежных оценок в регрессионную модель не следует включать слишком много факторов – их число не должно превышать одной трети объема имеющихся данных.

Практически множественный регрессионный анализ выполняется аналогично случаю парной линейной регрессии с учетом того, что в качестве независимой (объясняющей, экзогенной) переменной выбран не один, а несколько (множество) факторов (показателей). При выделении входного интервала *X,* помечаются столбцы значений всех независимых переменных вместе с названиями.

Использование фиктивных (бинарных) переменных в регрессионном анализе

До сих пор в качестве факторов мы рассматривали экономические переменные, принимающие количественные значения в некотором интервале. Однако при изучении связей между показателями результирующий признак, являющийся количественно переменной, может зависеть не только от количественных, но и от неколичественных (качественных) факторных признаков. Теория не накладывает никаких ограничений на характер регрессоров.

Переменные, входящие в состав регрессионной модели, могут принимать как конечное, так и бесконечное множество значений. Очевидно, что для включения неколичественной переменной в регрессионную модель необходимо каким-то образом поставить в соответствие ее качественным значениям

числовые величины. Это можно сделать с помощью фиктивных переменных (*dummy variables*).

Фиктивные переменные – это переменные бинарного типа, т.е. каждая переменная может принимать всего два значения: единица или ноль.

Бинарные переменные, несмотря на свою внешнюю простоту, являются весьма гибким инструментом при проведении различных эконометрических исследований.

Учет качественных признаков с помощью фиктивных переменных

В некоторых случаях отдельные факторы, которые мы хотели бы ввести в регрессионную модель, являются качественными по своей природе и, следовательно, не измеряются в числовой шкале.

Например, исследуется зависимость между возрастом и заработной платой населения по выборке, в которой представлены данные по лицам как мужского, так и женского пола. Необходимо выяснить, обусловлены ли различия в заработной плате полом индивидуума.

В принципе, возможным решением было бы оценивание отдельных регрессий для двух указанных категорий с последующим выяснением, различаются ли полученные коэффициенты. Другой возможный подход состоит в том, что мы оцениваем единую регрессию, с использованием всей совокупности наблюдений, но измеряем степень влияния качественного фактора при помощи введения фиктивной переменной.

Второй подход обладает двумя важными преимуществами: во-первых, имеется простой способ проверки, является ли воздействие качественного (не количественного) фактора значимым; во-вторых, при условии выполнения определенных предположений регрессионные оценки оказываются более эффективными.

Для исследования влияния качественных признаков в регрессионную модель вводятся фиктивные, или бинарные, переменные, которые, как правило,

принимают значение 1, если данный качественный признак присутствует в наблюдении, и значение 0 при его отсутствии. Как правило, фиктивная переменная обозначается буквой d.

Фиктивная переменная d – такая же «равноправная» переменная, как и любая другая экзогенная переменная (x). Ее «фиктивность» состоит только в том, что она количественным образом описывает качественный признак.