МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РФ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«Чеченский государственный университет»

Магомадов Э.М., Шабазов Р.Р., Янгульбаева Л.Ш.

ВВЕДЕНИЕ В ЭКОНОМЕТРИКУ

Учебное пособие

Грозный - 2015

Печатается по решению Ученого Совета
ФГБОУ ВО «Чеченский государственный университет»
Протокол № 11 от 31.12. 2015 г.

УДК 330.43(075.8)
ББК 65в6я73 К29
М-12

Рецензенты:

Авторханов А.И. – зав. кафедрой «Экономический анализ», доктор экономических наук, профессор (ФГБОУ ВО Чеченский государственный университет)

Мартынова М.А. – кандидат экономических наук, доцент (ФГБОУ ВО Грозненский государственный нефтяной технический университет)

Магомадов Э.М., Шабазов Р.Р., Янгульбаева Л.Ш. Введение в эконометрику. Учебное пособие. – Грозный: издательство ФГБОУ ВО «Чеченский государственный университет». 2015. – 88 с.

Пособие содержит краткое изложение основ эконометрики и написан на основе лекций, которые авторы в течение ряда лет читали в Чеченском государственном университете. Подробно изучаются линейные регрессионные модели (метод наименьших квадратов, проверка гипотез, автокорреляция ошибок, спецификация модели). В заключительной главе рассматриваются регрессионные модели с фиктивной переменной. Отдельный параграф посвящен средствам MS Excel в множественном регрессионном анализе. Приводятся расчеты с помощью инструментов «Регрессия», «Пакет анализа».

Главная цель пособия – предоставление информации, которую можно использовать при самостоятельном освоении материала.

Для студентов, аспирантов и преподавателей вузов.

Оглавление

Введение

Развитие экономики, усложнение экономических процессов и повышение требований к принимаемым управленческим решениям в области макро- и микроэкономики потребовало более тщательного и объективного анализа реально протекающих процессов на основе привлечения современных математических статистических методов.

С другой стороны, проблема нарушения предпосылок классических статистических методов при решении реальных экономических задач привела к необходимости развития и совершенствования классических методов математической статистики уточнения постановок соответствующих задач.

В результате этих процессов осуществилось выделение и формирование новой отрасли знания под названием «Эконометрика», связанной с разработкой и применением методов количественной оценки экономических явлений, процессов и их взаимосвязей.

Основным методом исследования в эконометрике является экономико-математическое моделирование. Правильно построенная модель должна давать ответ на вопрос о количественной оценке величины изменения изучаемого явления или процесса в зависимости от изменений внешней среды. Например, как скажется увеличение или уменьшение уровня инвестиций на совокупном валовом продукте, какие дополнительные ресурсы понадобятся для запланированного увеличения выпуска продукции и т.п.

Практическая значимость эконометрики определяется тем, что применение ее методов позволяет выявить реально существующие связи между явлениями, дать обоснованный прогноз развития явления в заданных условиях, проверить и численно оценить экономические последствия принимаемых управленческих решений.

Построение эконометрических моделей приходится осуществлять в условиях, когда нарушаются предпосылки классических статистических методов, и учитывать наличие таких явлений, как:

– мультиколлинеарность объясняющих переменных;

– закрытость механизма связи между переменными в изолированной регрессии;

– эффект гетероскедастичности, т. е. отсутствия нормального распределения остатков для регрессионной функции;

– автокорреляция остатков;

– ложная корреляция.

Разработка методов, преодолевающих эти трудности, составляет теоретическую основу эконометрики. Наряду с логически правильным формальным применением имеющегося математического и статистического инструментария важными составляющими успеха эконометрического исследования являются экономически адекватная постановка задачи и последующая экономическая интерпретация полученных результатов.

Огромный толчок развитию эконометрических методов и их широкому внедрению в практику дало развитие средств вычислительной техники и, особенно, появление персональных и портативных компьютеров. Разработка программных пакетов, реализующих методы построения и исследования эконометрических моделей привела к тому, что выполнение эконометрических процедур становится доступным самому широкому кругу аналитиков, экономистов и менеджеров. В настоящее время основные усилия прикладного исследователя сводятся к подготовке качественных исходных данных, к правильной постановке проблемы и экономически обоснованной интерпретации результатов исследования. Вместе с тем, от исследователя требуется четкое понимание областей применимости используемых методов и сложности и неочевидности процесса перенесения полученных теоретических результатов на реальную действительность. Настоящее пособие отражает содержание семестрового курса лекций, читаемых на факультете экономики и финансов студентам специальности «Бухгалтерский учет, анализ и аудит», «Экономическая теория», «Налоги и налогообложение», «Экономика и управление на предприятиях АПК», «Финансы и кредит», и соответствует Государственному образовательному стандарту по дисциплине «Эконометрика».

Изложение материала ориентировано на читателя, обладающего знаниями в пределах курсов высшей математики и теории вероятностей и математической статистики, читаемых студентам экономических специальностей. Пособие будет также полезно всем желающим познакомиться с основными задачами, моделями и методами эконометрики.

Глава 1. Эконометрика и эконометрическое моделирование: основные понятия и определения

1.1. Предмет и методы эконометрики

- Основная задача эконометрики – наполнить эмпирическим содержанием априорные экономические рассуждения.

Л. Клейн

- Эконометрика есть единство трёх составляющих – статистики, экономической теории и математики.

Р. Фриш

Эконометрика – это наука, которая даёт количественное выражение взаимосвязей экономических явлений и процессов.

Основные задачи эконометрики: *построение количественно определённых экономико-математических моделей, разработка методов оценки их параметров по статистическим данным, анализ свойств построенных моделей и прогнозирование на их основе экономических процессов.*

Можно выделить три основных класса моделей, которые применяются для анализа и прогнозирования экономических процессов:

- *модели временных рядов;*
- *регрессионные модели с одним уравнением;*
- *системы одновременных уравнений.*

При этом все переменные любой эконометрической модели по способу их вхождения в эту модель можно разбить на ***объясняемые (зависимые, исследуемые)*** переменные и ***объясняющие (предопределённые, факторные)*** переменные.

Например, если мы будем решать задачу прогнозирования продаж мороженого в определённый день каким-либо торговым предприятием, то объясняемой переменной будет объём продаж, а объясняющими переменными могут выступать: температура воздуха, торговая наценка, среднедушевой доход населения и другие.

Необходимым условием использования той или иной переменной при построении модели является наличие ряда данных наблюдений (измерений) величины этой переменной, либо получение ряда значений с использованием дополнительных вычислений на основе наблюдений о показателях, объясняющих интересующую нас переменную.

Например, определение достоверных значений среднедушевого дохода непосредственно по результатам опросов и бухгалтерской

отчётности может оказаться сложнее оценки изменения дохода на основе информации об изменении розничного оборота товаров и услуг, а также изменении общей суммы банковских вкладов населения.

В эконометрике выделяют три типа данных:

I. ***Кросс секционные (перекрёстные)*** данные представляют ситуацию в группе переменных в отдельный момент времени. Таковыми, например, являются публикуемые в деловых разделах газет списки цен на различные акции, процентные ставки по разным видам вкладов и обменные курсы разных валют. Другим примером может служить информация о продажах торговым предприятием в определённый день товаров различных групп (пищевых, хозяйственных и т.д.)

II. ***Пространственные*** данные характеризуют ситуацию по конкретной переменной (или набору переменных), относящейся к пространственно-разделённым однотипным объектам в один момент времени. Например, данные о курсах валют в один день по разным обменным пунктам города или продажи мороженого в различных киосках в один день.

III. ***Временные ряды*** отражают изменения (динамику) какой-либо переменной на промежутке времени. Например, данные об обменном курсе валюты за каждый день в конкретном обменном пункте или данные о продажах мороженого в одном киоске за каждый день будут являться ежедневным временным рядом.

Практическая значимость эконометрики определяется тем, что применение ее методов позволяет выявить реально существующие связи между явлениями, дать обоснованный прогноз развития явления в заданных условиях, проверить и численно оценить экономические последствия принимаемых управленческих решений.

Экономические явления взаимосвязаны и взаимообусловлены. Следствием этого является то, что значения соответствующих экономических показателей изменяются во времени с учетом этих взаимосвязей. Так, например, известно, что совокупный спрос зависит от уровня цен, потребление – от располагаемого дохода, инвестиции – от процентной ставки и так далее. Перед исследователем стоит задача выявления таких связей, их количественной оценки и изучения возможности использования выявленных связей в экономическом анализе и прогнозировании. Разработкой соответствующего инструментария и его применением

для решения конкретных практических экономических задач как раз и занимается эконометрика. В основе любого эконометрического исследования лежит построение экономико-математической модели, адекватной изучаемым реальным экономическим явлениям и процессам. Процесс построения эконометрических моделей начинается с качественного исследования проблемы методами экономической теории, далее формулируются цели исследования, выделяются факторы, влияющие на изучаемый показатель, и формулируются предположения о характере предполагаемой зависимости. На этой основе изучаемые зависимости выражаются в виде математических формул и соотношений. Следует отметить, что ввиду невозможности одновременно учесть большое количество факторов, влияющих на изучаемый показатель, предполагаемые зависимости между переменными будут выполняться не точно, а с определенной погрешностью. Кроме того, экономическим явлениям присуща внутренняя неопределенность, связанная с целенаправленной деятельностью субъектов экономики.

Вышесказанное обуславливает применение статистических методов, с помощью которых осуществляется отбор значимых факторов, определяется наличие и степень тесноты связи между изучаемыми показателями, дается количественная оценка параметров предполагаемых зависимостей и исследуется степень их соответствия реальной действительности. Основным инструментом математической статистики, используемым для построения эконометрических моделей, являются методы корреляционного и регрессионного анализа.

Корреляционный анализ ставит своей целью проверку наличия и значимости линейной зависимости между переменными без разделения переменных на зависимые и объясняющие. Ответ на эти вопросы дается с помощью вычисления показателей (коэффициентов) корреляции.

Регрессионный анализ направлен на выражение изучаемой зависимости в виде аналитической формулы с предварительным выделением зависимых и объясняющих переменных. Регрессионный анализ призван:

– выявить переменные, которые определяют поведение других величин и, следовательно, могут использоваться как объясняющие переменные;

– определить формулу зависимости и экономический смысл ее

коэффициентов.

Результатом проведения регрессионного анализа является построение, так называемого уравнения регрессии. После построения уравнения регрессии осуществляется проверка его статистического качества, включающая:

– проверку статистической значимости коэффициентов уравнения регрессии; – проверку общего качества уравнения регрессии;

– проверку наличия свойств данных, предполагавшихся при оценивании уравнения регрессии.

1.2. Соотношения между экономическими переменными

Одна из наиболее общих задач в экономических исследованиях состоит в оценивании степени зависимости изучаемой величины Y от одной или нескольких случайных (или неслучайных) величин X, называемых *факторами*. Зависимость может быть *функциональной, статистической*, либо отсутствовать вовсе.

Строгая функциональная зависимость между экономическими показателями (наличие всегда выполняющегося равенства Y=f(X)) реализуется редко, так как они подвержены влиянию случайных факторов. При статистической зависимости изменение одной из величин влечет изменение распределения другой (в частности, среднего значения; в этом случае статистическую зависимость называют *корреляционной*).

Причем, всегда есть несколько величин, которые определяют главные тенденции изменения рассматриваемой величины, и в экономической теории и практике ограничиваются тем или иным кругом таких величин (*объясняющих переменных*). Однако всегда существует и воздействие большого числа других, менее важных или трудно идентифицируемых факторов, приводящее к отклонению значений *объясняемой* (зависимой) переменной от конкретной формулы ее связи с объясняющими переменными, сколь бы точной эта формула ни была. Нахождение, оценка и анализ таких связей, идентификация объясняющих переменных, построение формул зависимости и оценка их параметров и составляют предмет **корреляционно-регрессионного анализа**, при этом *корреляционный анализ* занимается исследованием взаимозависимости случайных величин, тогда как *регрессионный анализ* на базе выборочных данных исследует зависимость случайной величины от ряда неслучайных и

случайных величин.

Примерами корреляционно, но не функционально, связанных величин являются объемы производства и себестоимость продукции, объемы продаж и прибыль, урожай зерна и количество внесенных удобрений. Действительно, в последнем примере с одинаковых по площади участков земли при равных количествах внесенных удобрений снимают различный урожай, т.е. отсутствует функциональная связь. Это объясняется влиянием случайных факторов (осадки, температура, качество семян и др.). Вместе с тем, как показывает опыт, *средний* урожай меняется с изменением количества удобрений, т.е. прослеживается корреляционная зависимость.

Рассмотрим сначала *однофакторную регрессионную модель*.

В этом случае имеется n пар наблюдений (x_i,y_i), i=1,2,…,n, над некоторыми случайными величинами X=$\{x_i\}$ и Y=$\{y_i\}$. Эти наблюдения можно представить точками на плоскости с координатами (x_i,y_i), получая так называемую *диаграмму рассеяния*. Задача построения регрессионной модели заключается в том, что необходимо подобрать некоторую кривую (график соответствующей функции) таким образом, чтобы она располагалась как можно “ближе” к этим точкам. Такого рода кривую называют *эмпирической* или *аппроксимирующей* кривой. Весьма часто тип эмпирической кривой определяется экспериментальными или теоретическими соображениями (исходя из законов экономической теории), в противном случае осуществить выбор кривой довольно трудно. Иногда точки на диаграмме рассеяния располагаются таким образом, что не наблюдается никакого их группирования, и, соответственно, нет никаких оснований предполагать наличие в наблюдениях какой-либо взаимозависимости.

Таким образом, результатом исследования статистической взаимозависимости на основе выборочных данных является построение *уравнений регрессии* вида $y=f(x)$.

1.3. Основные этапы построения эконометрической модели

Эконометрическое моделирование состоит из следующих этапов:

1. *Постановочный этап.* Формулируются конечные цели моделирования, определяются наборы возможных исследуемых (объясняемых) переменных $\overline{Y}=(y_1,y_2,\dots y_k)$ и факторных

(объясняющих) переменных $\overline{X} = (x_1, x_2, \ldots x_m)$.

2. *Предварительный этап.* Осуществляется предварительный анализ экономической сути изучаемого явления, возможностей сбора и обработки статистических данных.

3. *Этап параметризации.* Производится *выбор общего вида модели*, в том числе состава и формы входящих в неё связей. Например, может быть выбрана модель с одной объясняющей и одной объясняемой переменными – ***модель парной регрессии***. Если объясняющих (факторных) переменных используется две или более, то говорят об использовании ***модели множественной регрессии***. При этом, в качестве вариантов могут быть выбраны линейная, экспоненциальная, гиперболическая, показательная и другие виды функций, связывающие эти переменные.

4. *Информационный этап.* Сбор информации (проведение наблюдений, использование материалов отчётности и т.д.) и предварительный анализ данных (проверка аномальных значений показателей, сглаживание, тестирование на наличие тенденции исследуемых показателей к изменению).

5. *Идентификация модели.* Определению неизвестных параметров (коэффициентов) модели с использованием имеющегося набора данных. Наибольшее распространение для оценки параметров получил метод наименьших квадратов.

6. *Проверка (верификация) модели и прогнозирование.* Предполагает сопоставление реальных и модельных данных, проверку адекватности модели, оценку точности модельных данных. Если модель адекватна и имеет приемлемую точность, то на её основе строится прогноз – точечный и интервальный.

Контрольные вопросы

1. Охарактеризуйте предмет эконометрики.
2. Укажите основные этапы эконометрического исследования.
3. Какие задачи решают корреляционный и регрессионный анализы?
4. Каковы особенности причинно-следственных отношений в социально- экономических явлениях?
5. Какие зависимости называются стохастическими?
6. Какие типы данных используются в эконометрическом исследовании?
7. Опишите основные этапы построения эконометрической

модели.

8. Какие виды аналитических зависимостей, наиболее часто используются при построении моделей?

9. Какие методы используются для отбора факторов?

10. Какие методы используются для оценки параметров модели?

11. Какими свойствами характеризуется качество оценок параметров?

Глава 2. Парная регрессия и корреляция

Парная регрессия характеризует регрессию между двумя изучаемыми наблюдениями – y и x, иначе говоря – это модель вида:

$$y = \hat{f}(x),$$

где y – объясняемая (эндогенная) переменная (или же результативный признак); x – объясняющая (экзогенная) или независимая переменная (иначе признак-фактор). Обозначение «^» говорит о том, что между наблюдениями x и y нет строгой функциональной зависимости, а значит почти в каждом отдельном случае значение y есть сумма двух слагаемых:

$$y = y_x + \varepsilon,$$

где y – фактическое значение результативного признака; y_x – теоретически полученное значение результативного признака, которое получено исходя из уравнения регрессии; ε – ненаблюдаемая величина, которая характеризует отклонения фактических данных результативного признака от теоретических полученных из уравнения регрессии.

Также случайную величину ε называют возмущением. Она содержит влияние факторов, которые не учтены в модели, случайных ошибок и особенностей, связанных с измерениями. Ее присутствие в модели может быть связанно с тремя факторами:

- спецификацией модели;
- выборочным характером исходных данных;
- особенностями измерения переменных.

От правильно подобранной спецификации модели будет зависеть влияние случайных ошибок на построенную модель: чем ближе теоретически полученные значения y_x к значениям фактических данных y, тем меньше влияние случайных ошибок.

Также к ошибкам спецификации можно отнести неправильный выбор математической функции для y_x и недоучет в уравнении регрессии какого-либо существенного фактора, т.е. использование парной регрессии вместо множественной.

Помимо ошибок спецификации на модель могут оказывать влияние ошибки выборки, которые появляются в случае

неоднородности данных в исходной статистической совокупности, что, обычно бывает при анализе экономических процессов. В случае неоднородности изучаемой совокупности в построенном уравнении регрессии отсутствует всякий практический смысл. Чтобы получить хороший результат чаще всего исключают из совокупности единицы с аномальными значениями исследуемых признаков. В таком случае результаты построенной модели регрессии будут представлять из себя выборочные характеристики.

Использование временной информации также представляет собой выборку из всего множества хронологических дат. Изменив временной интервал, можно получить другие результаты регрессии.

Наибольшую опасность (опасения, трудности) в практическом применении регрессионных методов будут представлять ошибки измерения. Если ошибки спецификации можно снизить, изменив форму модели (вид математической формулы), а ошибки выборки – увеличив объем выборки, то ошибки измерения практически сводят к нулю все расчеты. Связанные с количественной оценкой связи между изучаемыми признаками.

Особенно велико влияние ошибок измерения в случае исследования на макроуровне. Так, при анализе спроса и потребления в качестве объясняющей переменной широко применяется «доход на душу населения». Вместе с тем, статистическое измерение величины дохода сопряжено с рядом трудностей и не лишено возможных ошибок, например, в результате наличия скрытых доходов.

Предположив, что ошибки измерения были сведены к минимуму, основной упор в эконометрических исследованиях делается на устранение ошибок спецификации модели.

При наличии парной регрессии общий вид математической функции $y_x = f(x)$ можно определить тремя методами:

1) графический метод;

2) аналитический метод, т.е. опираясь на теорию изучаемой взаимосвязи;

3) экспериментальный метод.

Применение графического метода дает наглядное представление относительно зависимости между изучаемыми признаками.

При изучении зависимости между двумя признаками графический метод подбора вида уравнения регрессии достаточно нагляден. Он базируется на анализе поля корреляции. Основные типы кривых, используемые при количественной оценке связей,

представлены на рис. 1.1:

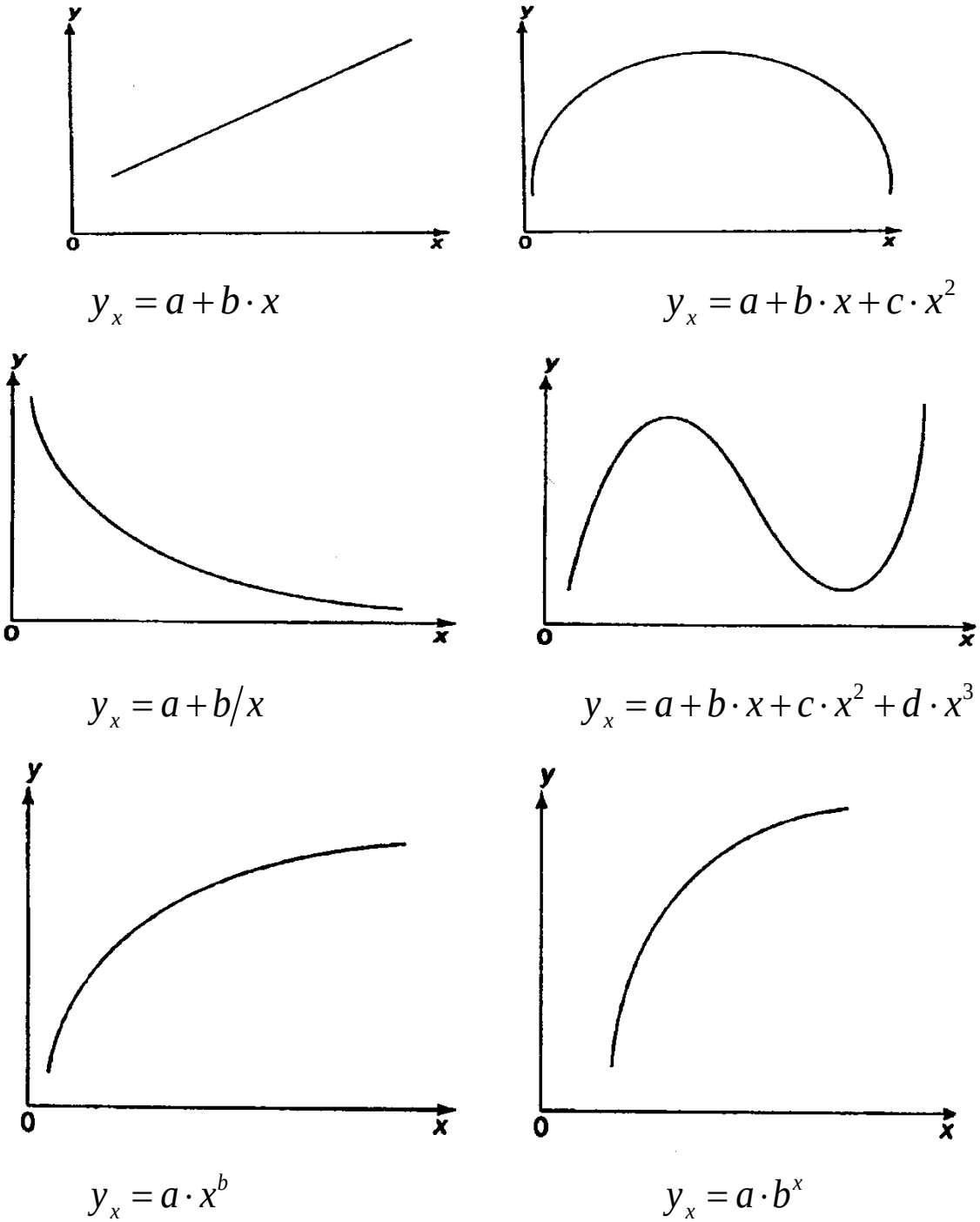


Рис. 1.1. Основные типы кривых, используемые при количественной оценке связей между двумя переменными.

Особый интерес при выборе типа уравнения регрессии для нас будет представлять аналитический метод. Он базируется на изучении материальной природы связи исследуемых признаков.

В случае анализа информации, основанном на использовании программных продуктов, вид уравнения регрессии выбирается экспериментальным методом

При обработке информации на компьютере выбор вида уравнения регрессии обычно осуществляется экспериментальным методом, т.е. иначе сравнивают величины остаточных дисперсий $\sigma^2_{\text{ост}}$, полученных из различных моделей регрессии.

Если уравнение регрессии проходит через все точки корреляционного поля, что возможно только при функциональной связи, когда все точки лежат на линии регрессии $y_x = f(x)$, то фактические значения результативного признака совпадают с теоретическими $y = y_x$, т.е. они полностью обусловлены влиянием фактора x. В этом случае остаточная дисперсия $\sigma^2_{\text{ост}} = 0$.

В практических исследованиях, как правило, имеет место некоторое рассеяние точек относительно линии регрессии.

Оно обусловлено влиянием прочих, не учитываемых в уравнении регрессии, факторов. Т.е., имеют место отклонения реальных данных от теоретически полученных $\left(y - y_x\right)$. Величина разностей фактических и теоретических данных $\left(y - y_x\right)$ является основой для расчета остаточной дисперсии:

$$\sigma^2_{\text{ост}} = \frac{1}{n}\sum\left(y - y_x\right)^2.$$

Отсюда можно сделать вывод, что для нас условием получения адекватных результатов анализа является близость остаточной дисперсии к нулю, ибо чем меньше остаточная дисперсия тем больше качество построенного уравнения регрессии, и соответственно меньше влияние факторов не включенных в наше уравнение

Считается, что число наблюдений должно в 7-8 раз превышать число рассчитываемых параметров при переменной x.Иначе говоря, строить модель линейной регрессии при менее, чем 7 наблюдениях, не имеет практического смысла. Если же вид парной линейной функции не подходит для анализа и её необходимо усложнить, тогда для анализа нам потребуется увеличить объем наблюдений, поскольку параметры x_i должны быть рассчитаны хотя бы по 7 наблюдениям. Т.е., в случае выбора параболы второй степени в качестве вида модели $\hat{y} = a + b \cdot x + c \cdot x^2$ отсюда следует, что информация должна быть уже по не менее, чем 14 наблюдениям.

2.1. Линейная модель парной регрессии и корреляции

Рассмотрим самый простой вид модели парной регрессии – линейную регрессию. Линейная регрессия широко применяется в эконометрике ввиду того что имеет четкую экономическую интерпретацию ее параметров.

Построение модели линейной регрессии сводится к нахождению уравнения вида

$$y_x = a + b \cdot x \text{ или } y = a + b \cdot x + \varepsilon. \qquad (1.1)$$

Уравнение вида $y_x = a + b \cdot x$ позволяет по заданным значениям фактора x находить теоретические значения результативного признака, подставляя в него фактические значения фактора x.

Весь процесс построения регрессии сводится к нахождению оценок коэффициентов регрессии – a и b

Наиболее популярным методом оценивания параметров регрессии является метод наименьших квадратов (МНК).

Применив МНК мы получим оценки параметров a и b, которые минимизируют сумму квадратов отклонений фактических значений результативного признака y от теоретических y_x:

$$\sum_{i=1}^{n}\left(y_i - y_{x_i}\right)^2 = \sum_{i=1}^{n}\varepsilon_i^2 \;\rightarrow\; \min. \qquad (1.2)$$

Т.е. из всего множества линий линия регрессии на графике выбирается так, чтобы сумма квадратов расстояний по вертикали между точками и этой линией была бы минимальной (рис. 1.2):

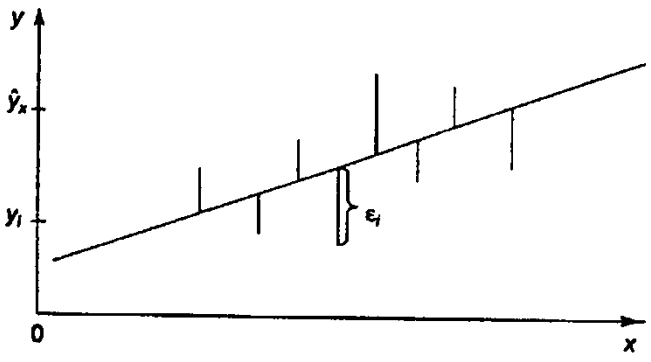


Рис. 1.2. Линия регрессии с минимальной дисперсией остатков.

Мы знаем, что по второй теореме Вейерштрасса функция достигает своего максимума или минимума в точке, в которой её производная равна нулю, значит, чтобы найти a и b нужно найти их

частные производные и приравнять и к нулю. Обозначив $\sum_i \varepsilon_i^2$ через S получим:

$$S = \sum(y_i - \hat{y}_i)^2 = \sum(y - a - b \cdot x)^2$$

$$\begin{cases} \frac{\partial S}{\partial a} = -2\Sigma(y - a - b \cdot x) = 0; \\ \frac{\partial S}{\partial b} = -2\Sigma(y - a - b \cdot x)\, x = 0. \end{cases}$$

(1.3)

$$\begin{cases} -2\sum y + 2\sum a + 2\sum bx = 0 \\ -2\sum yx + 2\sum ax + 2\sum bx^2 = 0 \end{cases}$$

$$\begin{cases} -\sum y + \sum a + \sum bx = 0 \\ -\sum yx + \sum ax + \sum bx^2 = 0 \end{cases}$$

$$\begin{cases} \sum a + \sum bx = \sum y \\ \sum ax + \sum bx^2 = \sum yx \end{cases}$$

Получили следующую систему линейных уравнений для оценки параметров a и b:

$$\begin{cases} na + b\sum x = \sum y \\ a\sum x + b\sum x^2 = \sum yx \end{cases} \quad (1.4)$$

Решив систему нормальных уравнений, мы получим интересующие нас коэффициенты a и b. Систему уравнений решим, либо путем последовательного исключения переменных, либо методом Кремера (метод определителей): $a = \frac{\Delta a}{\Delta}$; $b = \frac{\Delta b}{\Delta}$.

Также для нахождения параметров регрессии можно использовать следующие формулы:

$$a = \overline{y} - b \cdot \overline{x}, \qquad b = \frac{\mathrm{cov}(x, y)}{\sigma_x^2}, \quad (1.5)$$

Формула для отыскания параметра a выводится путем деления членов первого уравнения нашей системы на n.

$\mathrm{cov}(x,y) = \overline{y \cdot x} - \bar{y} \cdot \bar{x}$ – ковариация признаков x и y, $\sigma_x^2 = \overline{x^2} - \bar{x}^2$ – дисперсия признака x и

$$\bar{x} = \frac{1}{n}\sum x, \quad \bar{y} = \frac{1}{n}\sum y, \quad \overline{y \cdot x} = \frac{1}{n}\sum y \cdot x, \quad \overline{x^2} = \frac{1}{n}\sum x^2.$$

Ковариация – числовая характеристика совместного распределения двух случайных величин, которая равна математическому ожиданию произведения отклонений этих случайных величин от их математических ожиданий. Дисперсия – характеристика случайной величины, которая определяется как математическое ожидание квадрата отклонения случайной величины от ее математического ожидания. Математическое ожидание – представляет собой сумму произведений значений случайной величины на соответствующие вероятности.

Параметр b называют коэффициентом регрессии или же регрессором. Его величина характеризует среднее изменение *y* при изменении *x* на единицу своего измерения. Иначе говоря, отображает числовую зависимость переменной *y* от *x*.

Предположим, что функция спроса (y, тыс. руб.) выражена следующим образом: $\hat{y} = 278 - 1{,}4x$, (x – цена на товар, тыс. руб.). Значит с увеличением цены на товар на единицу стоимости спрос на этот товар будет снижаться в среднем на 1,4 тыс.руб.

Благодаря тому, что коэффициент регрессии b имеет четкую экономическую интерпретацию, линейное уравнение регрессии является распространенным в эконометрических исследованиях.

Параметр *a*, характеризует значение функции в точке, в которой переменная *x* равна нулю. Поскольку переменная *x* не может быть равна нулю, это значит, что *a* не может применяться для трактовки и описания экономических явлений.

Пример: по данным урожайности *y*(тонн) и затрат на минеральные удобрения *x*(тыс. руб.), приведенных в таблице (1.1.), построить модель, описывающую зависимость *y* от *x*.

Таблица 1.1.

№	***y***	***x***
1	10	5
2	12	6
3	15	8
4	16	9

5	18	10
6	21	11
7	24	14
8	26	16
Сумма	**142**	**79**

Чтобы построить модель, найдем все необходимые для системы (1.4) значения

№	y	x	yx	x^2
1	10	5	50	25
2	12	6	72	36
3	15	8	120	64
4	16	9	144	81
5	18	10	180	100
6	21	11	231	121
7	24	14	336	196
8	26	16	416	256
Сумма	**142**	**79**	**1549**	**879**

Получим систему нормальных уравнений:

$$\begin{cases} 8a + 79b = 142 \\ 79a + 879b = 1549 \end{cases}$$

Решим ее методом Кремера:

$$\Delta = \begin{vmatrix} 8 & 79 \\ 79 & 879 \end{vmatrix} = 791$$

$$\Delta a = \begin{vmatrix} 142 & 79 \\ 1549 & 879 \end{vmatrix} = 2447$$

$$\Delta b = \begin{vmatrix} 8 & 142 \\ 79 & 1549 \end{vmatrix} = 1174$$

$$a = \frac{\Delta a}{\Delta} = \frac{2447}{791} = 3.09$$

$$b = \frac{\Delta b}{\Delta} = \frac{1174}{791} = 1.48$$

Запишем уравнение регрессии: $\hat{y} = 3{,}09 + 1{,}48x$

Подставив в уравнение регрессии значения *x*, получим теоретические значения *y*:

$\hat{y}_1 = 3{,}09 + 1{,}48 \cdot 5 = 10{,}515;$

$\hat{y}_2 = 3{,}09 + 1{,}.48 \cdot 6 = 11{,}999;$

№	y	$\hat{y}$
1	10	10,515
2	12	11,999
3	15	14,967
4	16	16,451
5	18	17,936
6	21	19,42
7	24	23,872
8	26	26,841
Сумма	**142**	**142**

Полученные теоретические значения $\hat{y}$ близки к реальным значениям результата y, что дает нам возможность надеяться на то, что прогнозы, проведенные по данной модели, дадут хороший результат.

Коэффициент корреляции.

Коэффициент корреляции – R является показателем тесноты связи, построенное уравнение регрессии обычно дополняют им. В случае, если мы пользуемся линейной регрессией для анализа, тогда в качестве показателя тесноты связи будет выступать линейный коэффициент корреляцииR_{yx}, который можно рассчитать по следующей формуле:

$$R_{yx} = \frac{cov(x,y)}{\sigma_x \cdot \sigma_y} = \frac{\overline{yx} - \bar{y} \cdot \bar{x}}{\sqrt{(\overline{y^2} - \bar{y}^2) \cdot (\overline{x^2} - \bar{x}^2)}} \quad (1.6)$$

Линейный коэффициент корреляции меняется в пределах: $-1 \leq R_{yx} \leq 1$. Чем ближе значение коэффициента корреляции к ± 1, тем теснее связь между изучаемыми факторами (при $R_{yx} = \pm 1$будем иметь строгую функциональную зависимость). Чем ближе значение коэффициента корреляции к нулю, тем связь между факторами слабее.

Однако близость абсолютной величины линейного коэффициента корреляции еще не значит, что связь отсутствует вовсе, поскольку при другой (нелинейной) спецификации модели, возможно, что связь между признаками окажется достаточно тесной.

Воспользуемся табличными данными (1.1.) для расчета коэффициента корреляции.

y	x	yx	x^2	y^2
10	5	50	25	100
12	6	72	36	144
15	8	120	64	225
16	9	144	81	256
18	10	180	100	324
21	11	231	121	441
24	14	336	196	576
26	16	416	256	676
142	**79**	**1549**	**879**	**2742**
17,750	**9,875**	**193,625**	**109,875**	**342,750**

$$R_{yx} = \frac{193{,}625 - 17{,}750 \cdot 9{,}875}{\sqrt{(342{,}75 - 17{,}75^2) \cdot (109{,}875 - 9{,}875^2)}} = 0{,}99$$

Связь очень тесная, что говорит о сильной зависимости фактора *y* от *x*, в данном случае можно говорить о наличии функциональной связи.

Коэффициент детерминации.

Коэффициент детерминации позволяет оценить качество подобранной линейной функции, его рассчитывают как квадрат линейного коэффициента корреляции R_{yx}^2.

Коэффициент детерминации характеризует долю дисперсии результативного признака y, объясняемую регрессией, в общей дисперсии результативного признака:

$$\boldsymbol{R_{yx}^2 = 1 - \frac{\sigma_{\text{ост}}^2}{\sigma_y^2}}, \qquad (1.5)$$

где $\sigma_{\text{ост}}^2 = \frac{1}{n}\sum\left(y - y_x\right)^2$, $\sigma_y^2 = \frac{1}{n}\sum\left(y - \bar{y}\right)^2 = \overline{y^2} - \bar{y}^2$.

n	y	$\hat{y}$	$(y - \hat{y})^2$	$(y - \bar{y})^2$
1	10	10,515	0,265	60,063
2	12	11,999	0,000	33,063
3	15	14,967	0,001	7,563
4	16	16,451	0,204	3,063
5	18	17,936	0,004	0,063
6	21	19,42	2,497	10,563
7	24	23,872	0,016	39,063
8	26	26,841	0,707	68,063
Сумма	**142**	**142**	**3,694**	**221,500**

$$R_{yx}^2 = 1 - \frac{\sigma_{ост}^2}{\sigma_y^2} = 1 - \frac{3,694}{221,5} = 0,98$$

Полученный результат коэффициента детерминации говорит о том, что изменения результата уна 98% зависит от фактора *x*.

В случае наличия линейной связи между *y* и *x* коэффициент детерминации можно найти путем возведения в квадрат коэффициента корреляции.

Отсюда следует, что величина $1-r_{xy}^2$ будет характеризовать долю дисперсии y, обусловленную влиянием факторов, которые не были включены в модель.

После того, как найдено уравнение линейной регрессии, проводится оценка значимости как уравнения в целом, так и отдельных его параметров.

Проверить значимость уравнения регрессии – значит установить, соответствует ли математическая модель, выражающая зависимость между переменными, экспериментальным данным и достаточно ли включенных в уравнение объясняющих переменных (одной или нескольких) для описания зависимой переменной.

Ошибки аппроксимации.

Чтобы иметь представление относительно качества построенной модели из относительных отклонений по каждому наблюдению, определяют среднюю ошибку аппроксимации:

$$\bar{A} = \frac{1}{n}\sum\left|\frac{y - y_x}{y}\right| \cdot 100\%, \qquad (1.6)$$

где $\left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot 100\%$ – относительная ошибка аппроксимации, которая характеризует разницу между фактическими и теоретическими наблюдениями выраженную в процентах.

y_i	$\hat{y}$	$\left\|\frac{y_i - \hat{y}_i}{y_i}\right\| \cdot 100\%$
10	10,5145	5,1454
12	11,9987	0,0105
15	14,9671	0,2191
16	16,4513	2,8208
18	17,9355	0,3582
21	19,4197	7,5251
24	23,8723	0,5320
26	26,8407	3,2335
142	**142**	**19,8447**

Средняя ошибка аппроксимации не должна превышать 8–10%.

$$\bar{A} = \frac{1}{n} \cdot \sum \left| \frac{y_i - \hat{y}_i}{y_i} \right| \cdot 100\% = 2.48\%$$

Полученный результат не превышает допустимых приделов, что говорит о хорошем качестве построенной модели и соответственно данная модель применима для прогноза и анализа по ней поведения исследуемого показателя *y*.

2.2. Оценка значимости уравнения регрессии и его параметров

Оценка значимости уравнения регрессии.

Один из наиболее популярных вариантов оценки значимости уравнения регрессии заключается в следующем. Для построенного уравнения регрессии рассчитывается F – статистика, которая характеризует точность уравнения регрессии, данная статистика является отношением той части дисперсии объясняемой переменной, которая объясняется уравнением регрессии, к необъясненной (остаточной) части дисперсии.

Обычно дисперсионный анализ является предшествующим шагом к оценке значимости уравнения регрессии.

В математической статистике дисперсионный анализ рассматривается как самостоятельный инструмент статистического анализа. В эконометрике он применяется как вспомогательное средство для изучения качества регрессионной модели.

В дисперсионном анализе сумма квадратов разностей объясняемой переменной y от его среднего значения $\bar{y}$ можно разложить как две составляющие – сумму квадратов разностей, объясненную регрессией, и остаточную сумму квадратов разностей:

$$\sum (y - \bar{y})^2 = \sum \left(y_x - \bar{y}\right)^2 + \sum \left(y - y_x\right)^2,$$

где $\sum (y - \bar{y})^2$ – общая сумма квадратов разностей;

$\sum \left(y_x - \bar{y}\right)^2$ – сумма квадратов разностей, объясненная регрессией;

$\sum \left(y - y_x\right)^2$ –остаточная сумма квадратов разностей, которая характеризует влияние неучтенных в модели факторов.

Схема дисперсионного анализа имеет вид, представленный в

таблице 1.2 (n – число наблюдений, m – число параметров при переменной x).

Таблица 1.2

Компоненты дисперсии	Сумма квадратов	Число степеней свободы	Дисперсия на одну степень свободы
Общая	$\sum\left(y-\bar{y}\right)^2$	$n-1$	$S^2_{\text{общ}}=\dfrac{\sum\left(y-\bar{y}\right)^2}{n-1}$
Факторная	$\sum\left(y_x-\bar{y}\right)^2$	m	$S^2_{\text{факт}}=\dfrac{\sum\left(y_x-\bar{y}\right)^2}{m}$
Остаточная	$\sum\left(y-y_x\right)^2$	$n-m-1$	$S^2_{\text{ост}}=\dfrac{\sum\left(y-y_x\right)^2}{n-m-1}$

Определение дисперсии на одну степень свободы приводит дисперсии к сравнимому виду. Сопоставляя факторную и остаточную дисперсии в расчете на одну степень свободы, получим величину F-критерия Фишера:

$$F=\frac{S^2_{\text{факт}}}{S^2_{\text{ост}}}. \tag{1.7}$$

Фактическое значение F-критерия Фишера (1.7) сравнивается с табличным значением $F_{\text{табл}}\left(\alpha;k_1;k_2\right)$ при уровне значимости α и степенях свободы $k_1=m$ и $k_2=n-m-1$. При этом, если фактическое значение F-критерия больше табличного, то признается статистическая значимость уравнения в целом.

Для парной линейной регрессии $m=1$, поэтому

$$F=\frac{S^2_{\text{факт}}}{S^2_{\text{ост}}}=\frac{\sum\left(y_x-\bar{y}\right)^2}{\sum\left(y-y_x\right)^2}\cdot\left(n-2\right). \tag{1.8}$$

Величина F-критерия связана с коэффициентом детерминации r^2_{xy}, и ее можно рассчитать по следующей формуле:

$$F=\frac{r^2_{xy}}{1-r^2_{xy}}\cdot\left(n-2\right). \tag{1.9}$$

Оценка значимости коэффициентов регрессии.

Помимо оценки значимости уравнения регрессии в целом, также

на значимость в отдельности оцениваются его параметры.

Оценка значимости проводится по t – статистике Стьюдента.

С этой целью по каждому из параметров определяется его стандартная ошибка: m_b и m_a.

Стандартная ошибка коэффициента регрессии определяется по формуле:

$$m_b = \sqrt{\frac{S_{\text{ост}}^2}{\sum\left(x-\overline{x}\right)^2}} = \frac{S_{\text{ост}}}{\sigma_x \cdot \sqrt{n}}, \qquad (1.10)$$

где $S_{\text{ост}}^2 = \dfrac{\sum\left(y-y_x\right)^2}{n-2}$ – остаточная дисперсия на одну степень свободы.

Величина стандартной ошибки совместно с t-распределением Стьюдента при $n-2$ степенях свободы применяется для проверки существенности коэффициента регрессии и для расчета его доверительного интервала.

Оценку значимости коэффициента регрессии проводят путем сравнения его величины со стандартной ошибкой, иначе говоря определяют фактическое значение t-критерия Стьюдента: $t_b = \dfrac{b}{m_b}$.Фактическое значение t – критерия сравнивают с табличным значением t_T при определенном уровне значимости α и числе степеней свободы $(n-2)$.

m_b – стандартная ошибка регрессии, которая определяется по следующей формуле:

$$\boldsymbol{m_b = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2 \ /(n-2)}{\sum(x_i - \overline{x})^2}}}$$

Если $t_b > t_T$, принимается гипотеза H_1 о значимости коэффициента регрессии.

Если $t_b < t_{T,}$ принимается гипотеза H_0 о незначимости коэффициента регрессии.

Стандартную ошибку параметра a определим по формуле:

$$m_a = \sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{n-2} \cdot \frac{\sum x^2}{n\cdot\sum(x-\bar{x})^2}} \qquad (1.11)$$

Процедура оценки значимости параметра a совпадает с рассмотренной выше, для коэффициента регрессии. t-критерий для оценки значимости a: $t_a = \frac{a}{m_a}$, полученный результат сравниваем с табличным значением при том же уровне значимости и степени свободы $n-2$.

Если $t_a > t_T$ принимается гипотеза H_1 о значимости коэффициента регрессии.

Если $t_a < t_T$ принимается гипотеза H_0 о незначимости коэффициента регрессии.

Оценим значимость коэффициентов регрессии при уровне значимости $\alpha = 0{,}05$. Дополним таблицу необходимыми расчетами, чтобы найти стандартные ошибки для a и b:

№	y	x	x^2	$\hat{y}$	$\sum(y_i - \hat{y}_i)^2$	$\sum(x - \bar{x})^2$
1	10	5	25	10,5145	0,265	23,814
2	12	6	36	11,9987	0,000	15,054
3	15	8	64	14,9671	0,001	3,534
4	16	9	81	16,4513	0,204	0,774
5	18	10	100	17,9355	0,004	0,014
6	21	11	121	19,4197	2,497	1,254
7	24	14	196	23,8723	0,016	16,974
8	26	16	256	26,8407	0,707	37,454
Сумма	**142**	**79**	**879**	**142**	**3,694058**	**98,8752**

$$m_a = \sqrt{\frac{3{,}69}{8-2} \cdot \frac{879}{8 \cdot 98{,}875}} = 0{,}82$$

$$m_b = \sqrt{\frac{3{,}694/6}{98{,}875}} = 0{,}078$$

$$t_a = \frac{3{,}094}{0{,}82} = 3{,}74$$

$$t_b = \frac{1{,}48}{0{,}078} = 18{,}81$$

t_T =2,45 Значит, оба параметра регрессии a и b значимы, по ним принимается гипотеза H_1.

Доверительный интервал для коэффициента регрессии определяется как $b \pm t_{\text{табл}} \cdot m_b$. Поскольку знак коэффициента регрессии указывает на рост результативного признака y при увеличении признака-фактора x $(b > 0)$, уменьшение результативного признака при увеличении признака-фактора $(b < 0)$ или его независимость от независимой переменной $(b = 0)$ (см. рис. 1.3), то границы доверительного интервала для коэффициента регрессии не должны содержать противоречивых результатов, например, $-1{,}5 \le b \le 0{,}8$. Такого рода запись указывает, что истинное значение коэффициента регрессии одновременно содержит положительные и отрицательные величины и даже ноль, чего не может быть.

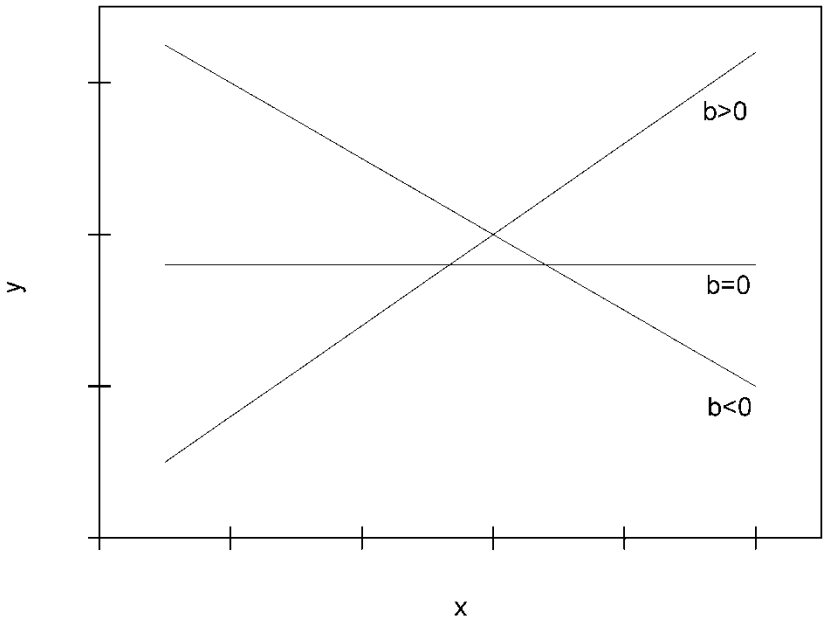


Рис. 1.3. Наклон линии регрессии в зависимости от значения параметра b.

Оценка значимости коэффициента корреляции – проводится по *t* - распределению Стьюдента.

С этой целью вычисляется t – фактическое по формуле:

$$t = \frac{R \cdot \sqrt{(n-2)}}{\sqrt{(1-R^2)}}$$

Если $t > t_T$, принимается гипотеза H_1 о значимости коэффициента корреляции $(R \neq 0)$.

Если $t < t_T$, принимается гипотеза H_0 о незначимости коэффициента корреляции $(R = 0)$.

$$t = \frac{0{,}99 \cdot \sqrt{6}}{\sqrt{(1 - 0{,}98)}} = 17{,}15$$

t – фактическое больше чем, $t_{\text{Т}}$, соответственно делаем вывод о значимости коэффициента корреляции.

Существует связь между t-критерием Стьюдента и F-критерием Фишера:

$$t_b = t_r = \sqrt{F}\,. \tag{1.12}$$

Точечный и интервальный прогноз.

В случае прогнозирования по уравнению регрессии вычисляется предсказываемое y_p значение как точечный прогноз y_x при $x_p = x_k$, т.е. путем подстановки в уравнение регрессии $y_x = a + b \cdot x$ соответствующего значения x, иначе говоря, суть точечного прогноза заключается в подстановке в полученную модель регрессии требуемого, прогнозного значения x. Однако ввиду того, что в парной линейной регрессии присутствует лишь один фактор x, точно определить будущее поведение изучаемого процесса не представляется возможным, поэтому точечный прогноз явно не реален. Поэтому его дополняют интервальным прогнозом, основанным на расчете стандартной ошибки y_p, т.е. m_{y_p}, и соответственно интервальной оценкой прогнозного значения y_p:

$$y_p - \Delta_{y_p} \le y_p \le y_p + \Delta_{y_p},$$

где $\Delta_{y_p} = m_{y_p} \cdot t_{\text{табл}}$, а m_{y_p} – средняя ошибка прогнозируемого индивидуального значения:

$$m_{y_p} = S_{\text{ост}} \cdot \sqrt{1 + \frac{1}{n} + \frac{\left(x_p - \overline{x}\right)^2}{n \cdot \sigma_x^2}}\,. \tag{1.13}$$

Рассмотрим другой пример.

Пример. Имеются данные о расходах населения на продукты питания и уровнях доходов, полученные в результате опроса, проведенного среди восьми групп семей.

Таблица 1.2

Расходы на продукты питания, y, тыс. руб.	0,9	1,2	1,8	2,2	2,6	2,9	3,3	3,8
Доходы семьи, x, тыс. руб.	1,2	3,1	5,3	7,4	9,6	11,8	14,5	18,7

Сделаем предположение о линейном характере связи между доходами семьи и расходами на продукты питания. Построим поле корреляции и убедимся в верности нашего предположения.

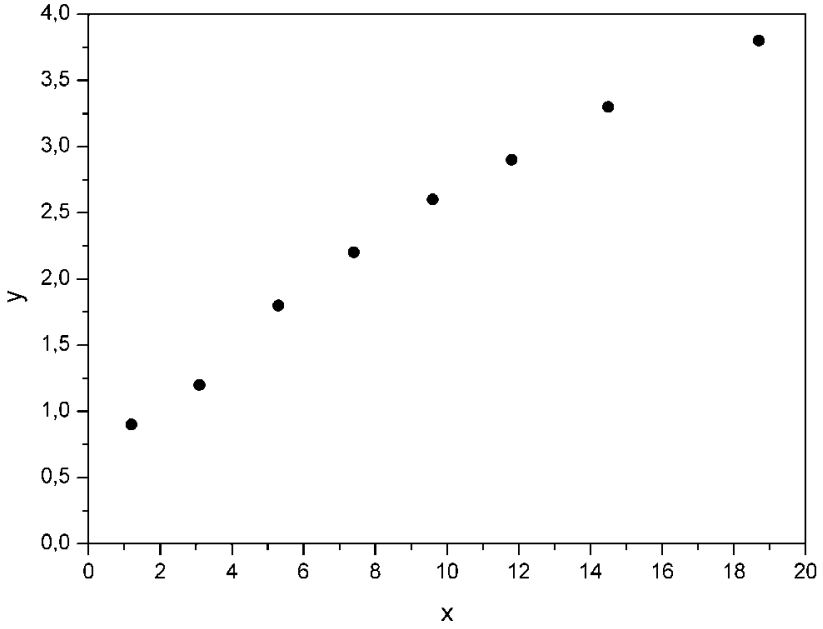


Рис. 1.4.

По графику видно, что точки выстраиваются в некоторую прямую линию.

Для удобства дальнейших вычислений составим таблицу.

Таблица 1.3

	x	y	yx	x^2	y^2	$\hat{y}_x$	$y-y_x$	$\left(y-y_x\right)^2$	A_i, %
1	1,1	1	1,1	1,21	1	1,0535	-0,0535	0,003	5,3534
2	3,2	1,3	4,16	10,24	1,69	1,4249	-0,125	0,016	9,6133
3	4,9	1,7	8,33	24,01	2,89	1,7256	-0,0257	0,001	1,5095
4	7,3	2,4	17,52	53,29	5,76	2,1501	0,24984	0,062	10,4099
5	9,4	2,5	23,5	88,36	6,25	2,5216	-0,0216	0,000	0,8641
6	11,9	3,1	36,89	141,61	9,61	2,9637	0,13621	0,019	4,3938
7	14,1	3,3	46,53	198,81	10,89	3,3529	-0,0529	0,003	1,6036
8	17,8	3,9	69,42	316,84	15,21	4,0073	-0,1074	0,012	2,7527
Итого	69,7	19,2	207,45	834,37	53,3	19,2	-0,0535	0,1149	36,500

Средн. Знач-е	8,7	2,4	25,931	104,296	6,663		–	0,014	4,56
σ	5,53	0,935	–	–	–	–	–	–	–
σ^2	30,56	0,874	–	–	–	–	–	–	–

Рассчитаем параметры линейного уравнения парной регрессии $y_x = a + b \cdot x$. Для этого воспользуемся формулами расчета коэффициентов регрессии:

$$b = \frac{cov(x,y)}{\sigma_x^2} = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{25{,}93 - 8{,}7 \cdot 2{,}4}{104{,}3 - 8{,}7^2} = 0{,}177$$

$$a = \bar{y} - b \cdot \bar{x} = 2{,}4 - 0{,}177 \cdot 8{,}7 = 0{,}859$$

В результате имеем модель вида: $\hat{y}_x = 0{,}859 + 0{,}177 \cdot x$. Делаем вывод, что при увеличении доходов семьи на 1000 руб. расходы на питание при этом увеличатся в среднем на 177 руб.

Как нами было отмечено ранее, уравнение линейной регрессии всегда дополняют коэффициентом корреляции, который характеризует тесноту связи между изучаемыми явлениями R_{xy}:

$$R_{yx} = \frac{cov(x,y)}{\sigma_x \cdot \sigma_y} = \frac{\overline{yx} - \bar{y} \cdot \bar{x}}{\sqrt{\left(\overline{y^2} - \bar{y}^2\right) \cdot \left(\overline{x^2} - \bar{x}^2\right)}}$$

$$= \frac{25{,}93 - 2{,}4 \cdot 8{,}7}{\sqrt{(6{,}7 - (2{,}4)^2) \cdot (104{,}3 - 8{,}7^2)}} = 0{,}99$$

Близость коэффициента корреляции к единице указывает на тесную линейную связь между признаками.

В результате коэффициент детерминации R^2 будет равен $0,98$. Коэффициент детерминации показывает, что построенное уравнение регрессии объясняет 98,4% дисперсии результативного признака, соответственно 1,6% приходится на долю случайных и неучтенных в модели факторов.

Для оценки качества построенной модели в целом воспользуемся F – критерием Фишера. Вычислим значение F – критерия для данного уравнения регрессии:

$$F = \frac{R_{xy}^2}{(1 - R_{xy}^2)} \cdot (n - 2) = \frac{0{,}984}{(1 - 0{,}984)} \cdot 6 = 371{,}002$$

Табличное значение $(k_1 = 1, \quad k_2 = n - 2 = 6, \quad \alpha = 0{,}05)$: $F_{\text{табл}} = 5{,}99$.

Поскольку $F_{\text{факт}} > F_T$, делаем вывод о том что уравнение регрессии в целом статистически значимо.

Чтобы проверить статистическую значимость коэффициентов регрессии по t – критерию Стьюдента, найдем их стандартные ошибки m_a, m_b.

$$m_b = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2 \ /(n-2)}{\sum(x_i - \bar{x})^2}} = 0{,}009$$

$$m_a = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2} \cdot \frac{\sum x^2}{n \cdot \sum(x - \bar{x})^2}} = 0{,}094$$

Вычислим t – статистики коэффициентов a и b:

$$t_a = \frac{0.859}{0.094} = 9{,}16$$

$$t_b = \frac{0.177}{0.009} = 19{,}26$$

При уровне значимости $\alpha = 0{,}05$ и числе степеней свободы $\nu = n - 2 = 6$ получим табличное значение критерия Стьюдента $t_{\text{Т}}$ = 2,447

Поскольку $t_b > t_{\text{табл}}$, $t_a > t_{\text{табл}}$ и $t_r > t_{\text{табл}}$, приходим к выводу, что параметры регрессии статистически значимы. Рассчитаем для коэффициентов регрессии a и b доверительные интервалы: $a \pm t \cdot m_a$ и $b \pm t \cdot m_b$. Получили, что $a \in [0{,}597;\ \ 1{,}075]$ и $b \in [0{,}145;\ \ 0{,}191]$.

Оценим статистическую значимость коэффициента корреляции по t – критерию Стьюдента:

$$t = \frac{R \cdot \sqrt{(n-2)}}{\sqrt{(1-R^2)}} = 19{,}26$$

Так как $t > t_{\text{Т}}$, делаем вывод о статистической значимости коэффициента корреляции.

Найдем среднюю ошибку аппроксимации (найдем ее с опираясь

на расчеты, приведенные в таблице). $A_i = \left| \frac{y_i - y_{x_i}}{y_i} \right| \cdot 100\% \, \overline{A} =$ 4,56%. Отклонение фактических наблюдений от теоретических составляет в среднем 4,56%, что говорит о хорошем качестве уравнения регрессии, т.е. свидетельствует о хорошем подборе модели к исходным данным. Данная модель адекватна и применима для прогноза по ней.

Спрогнозируем расходы на питание в случае когда доходы семьи будут составлять 9,85 тыс. руб., т.е. при условии, что признак – фактор составит 110% от среднего уровня:

$$\hat{y}_x = 0{,}859 + 0{,}177 \cdot 9{,}85 = 2{,}601 (\text{тыс. руб.})$$

При доходах семьи 9,845 тыс. руб. прогнозируемые расходы составят 2,601 тыс. руб. (следует помнить, что линейная модель имеет средний приближенный характер и соответственно опираться на один лишь точечный прогноз не стоит).

Дополним наш прогноз интервальным, т.е. найдем возможные интервалы варьирования прогноза. Для начала отыщем стандартную ошибку прогноза:

$$m_{\hat{y}_p} = \sqrt{\frac{\sum(y - \hat{y})^2}{n-2}} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{n \cdot \sum(x - \bar{x})^2}}$$

$$= \sqrt{\frac{0{,}115}{6}} \cdot \sqrt{1 + \frac{1}{8} + \frac{(9.845 - 8{,}713)^2}{8 \cdot 238{,}013}} = 0{,}138$$

$$\hat{y}_p - m_{\hat{y}_p} \cdot t_T \le \hat{y}_p \le \hat{y}_p + m_{\hat{y}_p} \cdot t_T$$

$2{,}113 < y_p < 2{,}867$.

Т.е. прогноз является статистически надежным.

Теперь на одном графике изобразим исходные данные и линию регрессии:

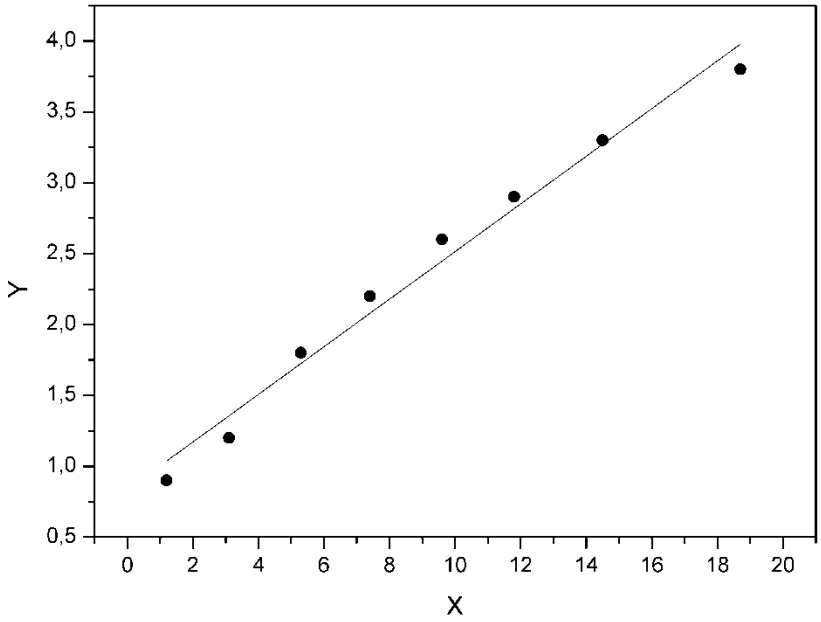


Рис. 1.5.

2.3. Построение парной линейной регрессии в Excel

Имеются данные в таблице, в которой приведены ежегодные значения денежной массы и национального дохода некоторой гипотетической страны (все величины выражены в миллиардах кварков (название национальной валюты)).

n	**Денежная масса (Y)**	**Национальный доход (X)**
1	1	1,1
2	1,3	3,2
3	1,7	4,9
4	2,4	7,3
5	2,5	9,4
6	3,1	11,9
7	3,3	14,1
8	3,9	17,8

Построить модель регрессии, характеризующей зависимость денежной массы от уровня национального дохода.

Заносим данные по задаче в таблицу Excel

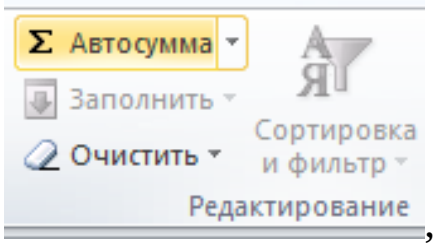
	A	B	C	D
1				
2				
3	n	y	x	
4	1	1	1,1	
5	2	1,3	3,2	
6	3	1,7	4,9	
7	4	2,4	7,3	
8	5	2,5	9,4	
9	6	3,1	11,9	
10	7	3,3	14,1	
11	8	3,9	17,8	

Для заполнения системы нормальных уравнений добавим столбцы: Y^2, X^2,XУ.

	A	B	C	D	E	F
1						
2						
3	n	Y	X	X^2	XY	Y^2
4	1	1	1,1			
5	2	1,3	3,2			
6	3	1,7	4,9			
7	4	2,4	7,3			
8	5	2,5	9,4			
9	6	3,1	11,9			
10	7	3,3	14,1			
11	8	3,9	17,8			

Расчет X^2 для первого наблюдения осуществляется путем ввода в ячейку D3 формулы «=C4^2». Последующее протягивание (указателем мыши за черную точку в правом нижнем углу ячейки D3)до ячейки D11. Наблюдения для $У^2$ найдем аналогичным способом. Расчет наблюдений ХУ осуществим путем ввода в E4 формулы «=B4*C4», с последующим протягиванием до ячейки E11.

Расчет сумм соответствующих столбцов можно выполнить различными способами, мы пойдем по наиболее легкому пути. Щелкнем в ячейке столбца B12 и щелкнем на значке автосуммы



, сразу будет выделен столбец «B4:B11», нажимаем Enter, в

выделенной ячейке появится значение суммы столбца, тем же протягиванием найдем суммы для столбцов X,X^2,XУ,$У^2$.

	A	B	C	D	E	F
1						
2						
3	n	**Y**	**X**	X^2	XY	Y^2
4	1	1	1,1	1,21	1,1	1
5	2	1,3	3,2	10,24	4,16	1,69
6	3	1,7	4,9	24,01	8,33	2,89
7	4	2,4	7,3	53,29	17,52	5,76
8	5	2,5	9,4	88,36	23,5	6,25
9	6	3,1	11,9	141,61	36,89	9,61
10	7	3,3	14,1	198,81	46,53	10,89
11	8	3,9	17,8	316,84	69,42	15,21
12	сумма	19,2	69,7	834,37	207,45	53,3

$$\begin{cases} 8a + b69{,}7 = 19{,}2 \\ 69{,}7a + b834{,}37 = 207{,}45 \end{cases}$$

Решая систему методом Крамера, запишем

8	3,9	17,8	316,84
сумма	19,2	69,7	834,37
дельта	=A11*D12-C12*C12		

Аналогично найдем $\Delta a, \Delta b$.

дельта	1816,87
дельта(a)	1560,64
дельта(b)	321,36

Расчитаем a как отношение $a = \frac{\Delta a}{\Delta}$ и b как $b = \frac{\Delta b}{\Delta}$.

В выделенной ячейке записываем расчет параметра a как «=B15/B14» и b«=B16/B14»

14	дельта	1816,87		
15	дельта(a)	1560,64	a	0,85897
16	дельта(b)	321,36	b	=B16/B14

Для удобства можно округлить содержимое ячеек D15 и D16, в которых у нас записаны вычисленные значения параметров aи b.

a	0,85897
b	0,17688

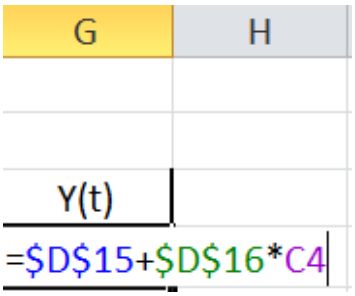
Для этого выделим их и вызовем контекстное меню, далее – «Формат ячеек – числовой – число десятичных знаков: 2 - Ок».

Соотвественно имеем модель типа $\hat{y} = 0,86 + 0,18x$.

Рассчитаем прогнозные (теоретические) значения Y для каждого имеющегося в таблице значения объясняющей переменной X.

Для этого введем в расчетной таблице дополнительный столбец G, озаглавив его «Y(t)».

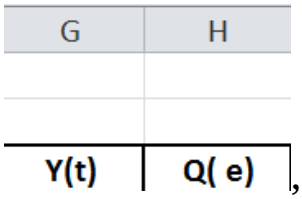
Т.к. уравнение регрессии уже получено, осталось лишь для каждого *i*-го значения X рассчитать в столбце G соответствующее прогнозное значение Y. Для этого в ячейке G4 введём формулу «=$D$15+$D$16*C4»



и протянем по диапазону ячеек G4:G11.

	A	B	C	D	E	F	G
1							
2							
3	n	Y	X	X^2	XY	Y^2	Y(t)
4	1	1	1,1	1,21	1,1	1	1,05
5	2	1,3	3,2	10,24	4,16	1,69	1,42
6	3	1,7	4,9	24,01	8,33	2,89	1,73
7	4	2,4	7,3	53,29	17,52	5,76	2,15
8	5	2,5	9,4	88,36	23,5	6,25	2,52
9	6	3,1	11,9	141,61	36,89	9,61	2,96
10	7	3,3	14,1	198,81	46,53	10,89	3,35
11	8	3,9	17,8	316,84	69,42	15,21	4,01
12	сумма	19,2	69,7	834,37	207,45	53,3	19,2

Найдем сумму квадратов отклонений между теоретическими и фактическими наблюдениями $\sum(y - \hat{y})^2$, добавим столбец в таблицу *Qe*(ε)



запишем в ячейку формулу «=(B4-G4)^2», заполним столбец:

n	Y	X	X^2	XY	Y^2	Y(t)	Q(e)
1	1	1,1	1,21	1,1	1	1,05	0,003
2	1,3	3,2	10,24	4,16	1,69	1,42	0,016
3	1,7	4,9	24,01	8,33	2,89	1,73	0,001
4	2,4	7,3	53,29	17,52	5,76	2,15	0,062
5	2,5	9,4	88,36	23,5	6,25	2,52	0,000
6	3,1	11,9	141,61	36,89	9,61	2,96	0,019
7	3,3	14,1	198,81	46,53	10,89	3,35	0,003
8	3,9	17,8	316,84	69,42	15,21	4,01	0,012
сумма	**19,2**	**69,7**	**834,37**	**207,45**	**53,3**	**19,2**	**0,115**

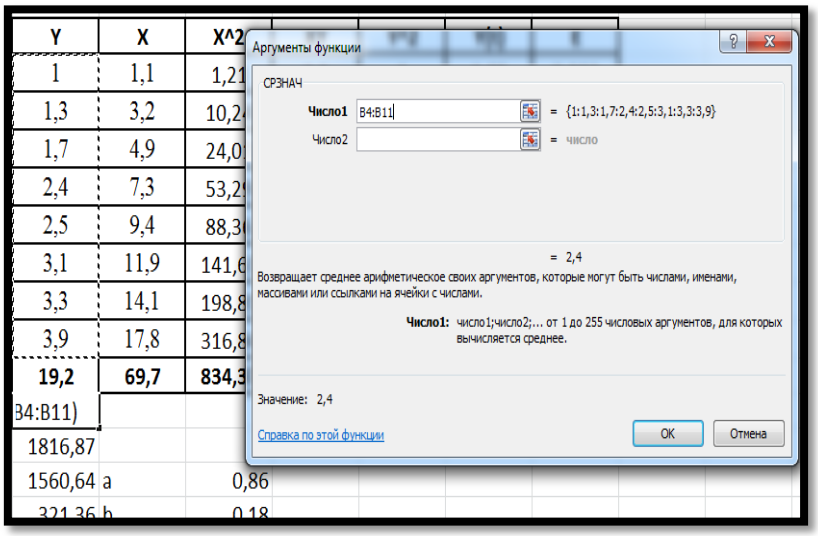
Далее расчитаем ***коэффициент корреляции.***

$$R_{yx} = \frac{cov(x,y)}{\sigma_x \cdot \sigma_y} = \frac{\overline{yx} - \bar{y} \cdot \bar{x}}{\sqrt{(\overline{y^2} - \bar{y}^2) \cdot (\overline{x^2} - \bar{x}^2)}}$$

Либо по формуле $R = \sqrt{1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2}}$. Обе формулы легко рассчитать в Excel.

В случае с первым подходом необходимо найти среднее арифмитическое значение для следующих столбцов: ***X, Y, XY, X^2, Y^2***.

Меню – «Формулы» - «Статистические» - «Ср. знач»



Вводим значения столбца Y – Ок., найдем средние аналогично и для остальных столбцов.

12	сумма	**19,2**	**69,7**	**834,37**	**207,45**	**53,3**	**19,2**	**0,115**
13	**среднее**	2,40	8,71	104,30	25,93	6,66	2,40	0,01

Далее получим результат, подставив в формулу соответствующие значения.

По второй формуле найдем значение суммы квадратов разностей

случайной величины и ее средней $\sum(\boldsymbol{y} - \bar{\boldsymbol{y}})^{\boldsymbol{2}}$.
Добавим столбец в таблицу, обозначим его через Q. Вводим в ячейку «=(B4-$B$13)^2»

n	Y	X	X^2	XY	Y^2	Y(t)	Q(e)	Q
1	1	1,1	1,21	1,1	1	1,05	0,003	=(B4-$B$13)^2
2	1,3	3,2	10,24	4,16	1,69	1,42	0,016	
3	1,7	4,9	24,01	8,33	2,89	1,73	0,001	
4	2,4	7,3	53,29	17,52	5,76	2,15	0,062	
5	2,5	9,4	88,36	23,5	6,25	2,52	0,000	
6	3,1	11,9	141,61	36,89	9,61	2,96	0,019	
7	3,3	14,1	198,81	46,53	10,89	3,35	0,003	
8	3,9	17,8	316,84	69,42	15,21	4,01	0,012	
сумма	**19,2**	**69,7**	**834,37**	**207,45**	**53,3**	**19,2**	**0,115**	
среднее	2,40	8,71	104,30	25,93	6,66	2,40	0,01	

	A	B	C	D	E	F	G	H	I
1									
2									
3	n	Y	X	X^2	XY	Y^2	Y(t)	Q(e)	Q
4	1	1	1,1	1,21	1,1	1	1,05	0,003	1,96
5	2	1,3	3,2	10,24	4,16	1,69	1,42	0,016	1,21
6	3	1,7	4,9	24,01	8,33	2,89	1,73	0,001	0,49
7	4	2,4	7,3	53,29	17,52	5,76	2,15	0,062	0
8	5	2,5	9,4	88,36	23,5	6,25	2,52	0,000	0,01
9	6	3,1	11,9	141,61	36,89	9,61	2,96	0,019	0,49
10	7	3,3	14,1	198,81	46,53	10,89	3,35	0,003	0,81
11	8	3,9	17,8	316,84	69,42	15,21	4,01	0,012	2,25
12	сумма	**19,2**	**69,7**	**834,37**	**207,45**	**53,3**	**19,2**	**0,115**	**7,220**
13	**среднее**	2,40	8,71	104,30	25,93	6,66	2,40	0,01	

Впишем в ячейку F15 комментарий «R= » , а в ячейку G15 формулу «=корень(1 – H12/I12)»

$$\boldsymbol{R} = \sqrt{\mathbf{1} - \frac{\mathbf{0,115}}{\mathbf{7,220}}} = \mathbf{0,99}$$

Коэффициент корреляции очень близок к единице, здесь можно говорить о прямой функциональной связи. Связь очень тесная прямая, соответственно – крайне сильная зависимость денежной массы от национального дохода.

Вычислим коэффициент детерминации как квадрат коэффициента корреляции: впишем в ячейку F16 комментарий «R^2 - = » , а в ячейку G16 формулу «= G15^2»
$R^2 = 0.98$ т .е. 98 % дисперсии y объеснено дисперсией $\hat{y}$.

Оценка значимости уравнения регрессии

Оценим значимость модели в целом. Или проще: «насколько модели можно доверять при имеющихся исходных данных».

Как известно, уравнение регрессии значимо, если наблюдаемое значение статистики F больше табличного значения F-критерия Фишера-Снедекора на уровне α (обычно α = 0,05) при k1 = p = m − 1 и k2 = n − m = n − p − 1 степенях свободы:

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

Впишем в ячейку F17 комментарий «F =» и в ячейку G17 формулу «=H13*(B14-2)/(I13-H13)».

$$\boldsymbol{F = \frac{0,98\,(8-2)}{(1-0,98)} = 371}$$

Полученное значение F = 371 надо сравнить с табличным значением. В таблице F-критерия Фишера для уровня значимости α = 0,05 выберем столбец $k1 = 1$ и строку $k2 = n – 2 = 8 – 2 = 6$. $F_T = 5,99$. Поскольку $F > F_T$

Примем гипотезу H_1 о том что полученная модель регрессии значима на уровне значимости α = 0,05 .

Ошибки аппроксимации

Фактические значения результативного признака отличаются от теоретических, рассчитанных по уравнению регрессии. Чем меньше эти отличия, тем ближе теоретические значения к эмпирическим данным, тем лучше качество модели. Величина отклонений фактических и расчетных значений результативного признака $(y - \hat{y})$ по каждому наблюдению представляет собой ошибку аппроксимации.

Поскольку $(y - \hat{y})$ может быть величиной как положительной, так и отрицательной, ошибки аппроксимации для каждого наблюдения принято определять в процентах по модулю.

$$A = \left|\frac{(y - \hat{y})}{y}\right| \cdot 100$$

Оценим качество полученного уравнения регрессии с помощью средней относительной ошибки аппроксимации.

Введем новый столбец, обозначим его «А», чтобы найти ошибки аппроксимации по рядам наблюдений:

	A	B	C	D	E	F	G	H	I	J
1										
2										
3	n	Y	X	X^2	XY	Y^2	Y(t)	Q(e)	Q	A
4	1	1	1,1	1,21	1,1	1	1,05	0,003	1,96	
5	2	1,3	3,2	10,24	4,16	1,69	1,42	0,016	1,21	
6	3	1,7	4,9	24,01	8,33	2,89	1,73	0,001	0,49	
7	4	2,4	7,3	53,29	17,52	5,76	2,15	0,062	0	
8	5	2,5	9,4	88,36	23,5	6,25	2,52	0,000	0,01	
9	6	3,1	11,9	141,61	36,89	9,61	2,96	0,019	0,49	
10	7	3,3	14,1	198,81	46,53	10,89	3,35	0,003	0,81	
11	8	3,9	17,8	316,84	69,42	15,21	4,01	0,012	2,25	
12	сумма	**19,2**	**69,7**	**834,37**	**207,45**	**53,3**	**19,2**	**0,115**	**7,220**	

В ячейку *J4* вводим формулу «= abs((B4 – G4)/B4)*100»,

Q	A
1,96	=ABS((B4-G4)/B4)*100
1,21	

Протянем формулу по диапазону «J4:J11» получим столбец отклонений фактических данных от теоретических в процентном выражении:

	A	B	C	D	E	F	G	H	I	J
1										
2										
3	n	Y	X	X^2	XY	Y^2	Y(t)	Q(e)	Q	A
4	1	1	1,1	1,21	1,1	1	1,05	0,003	1,96	5,35
5	2	1,3	3,2	10,24	4,16	1,69	1,42	0,016	1,21	9,61
6	3	1,7	4,9	24,01	8,33	2,89	1,73	0,001	0,49	1,51
7	4	2,4	7,3	53,29	17,52	5,76	2,15	0,062	0	10,41
8	5	2,5	9,4	88,36	23,5	6,25	2,52	0,000	0,01	0,86
9	6	3,1	11,9	141,61	36,89	9,61	2,96	0,019	0,49	4,39
10	7	3,3	14,1	198,81	46,53	10,89	3,35	0,003	0,81	1,60
11	8	3,9	17,8	316,84	69,42	15,21	4,01	0,012	2,25	2,75
12	сумма	**19,2**	**69,7**	**834,37**	**207,45**	**53,3**	**19,2**	**0,115**	**7,220**	**36,50**

Для получения средней ошибки аппроксимации в ячейку J 13 введем формулу «СРЗНАЧ(J4:J11)».

A
5,35
9,61
1,51
10,41
0,86
4,39
1,60
2,75
36,50
=срзнач(J4:J11)

В ячейке J13 получили значение$\overline{A} = 4{,}56\%$. Поскольку средняя ошибка аппроксимации не превышает пределов в 8 -10%, значит разброс наблюдаемых значений относительно оценочных мал, предполагаем, что данная модель достаточно точна и применима для прогноза, а так же можно надеяться на адекватное отражение зависимости изучаемого явления от случайной величины.

Коэффициент эластичности

Оценим силу связи между X и Y с помощью среднего коэффициента эластичности. Т.к. для парной линейной регрессии f '(x) = *b*, тогда

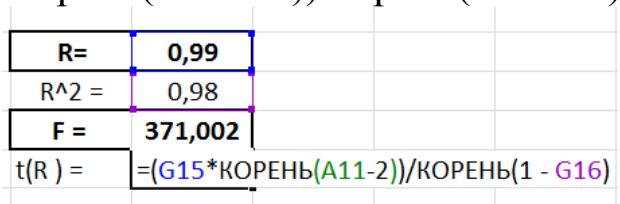
$$\overline{Э} = f'(x)\frac{\overline{x}}{\overline{y}} = b\frac{\overline{x}}{\overline{y}}$$

В ячейку I15 введем комментарий «$Э_{ср}$»,в ячейке J15 вводим формулу «=D15*(C13/B13)». Получим значение среднего коэффициента эластичности $\overline{Э} = 0{,}64$. Это означает, что при увеличении национального дохода на 1% от своего среднего значения, денежная масса увеличится на 0,64% от своего среднего значения. Сила влияния X на У не слишком велика.

Оценка значимости коэффициента корреляции

Оценка значимости коэффициента корреляции проводится по *t* – критерию Стьюдента. Для определения значимости по *t*-критерию воспользуемся формулой $t = \frac{R\sqrt{(n-2)}}{\sqrt{(1-R^2)}}$

В ячейке F18 введем комментарий «t (R) =», в ячейке G18 введем формулу « = (G15*корень(A11 – 2))/корень(1 – G16)» .



Отсюда *t* = 19,26

Сравниваем с t_T, которое определяется по таблице (критические значения *t*-критерия Стьюдента на уровне значимости 0,10; 0,05; 0,01) При уровне значимости α = 0,05 и степени свободы (*n* – 2), равной 6, табличное значение Стьюдента равно $t_{0,05(8-2)} = 2{,}45$

Число средней свободы *df*	α		
	0,10	0,05	0,01
1	6,3138	12,706	63,657
2	2,9200	4,3027	9,9248
3	2,3534	3,1825	5,8409
4	2,1318	2,7764	4,6041
5	2,0150	2,5706	4,0321
6	1,9432	2,4469	3,7074
7	1,8946	2,3646	3,4995
8	1,8595	2,3060	3,3554
9	1,8331	2,2622	3,2498

Получим: $t > t_T \rightarrow H_1$

Делаем предположение о значимости коэффициента корреляции.

Оценка значимости коэффициентов регрессии

Для оценки значимости коэффициентов модели регрессии воспользуемся следующими формулами:

$t_a = \frac{a}{m_a}$, где стандартная ошибка параметра *a* определяется как

$$m_a = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2} \cdot \frac{\sum x^2}{n \cdot \sum(x - \bar{x})^2}},$$

$t_b = \frac{b}{m_b}$ стандартная ошибка регрессора определяется по следующей формуле $m_b = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2 / (n-2)}{\sum(x_i - \bar{x})^2}}$

Для нахождения статистики параметров *a* и *b*введем в таблицу столбец $\sum(x_i - \bar{x})^2$, в ячейку «К3» введем комментарий Q(x), который представляет собой условное обозначение суммы квадратов разностей случайной величины x относительно ее математического ожидания.

J	K

A	Q(x)
5,35	
9,61	
1,51	
10,41	
0,86	
4,39	
1,60	
2,75	
36,50	

В ячеку «К4» запишем формулу «=(C4-$C$13)^2»

X	X^2	XY	Y^2	Y(t)	Q(e)	Q	A	Q(x)
1,1	1,21	1,1	1	1,05	0,003	1,96	5,35	=(C4-$C$13)^2
3,2	10,24	4,16	1,69	1,42	0,016	1,21	9,61	
4,9	24,01	8,33	2,89	1,73	0,001	0,49	1,51	
7,3	53,29	17,52	5,76	2,15	0,062	0	10,41	
9,4	88,36	23,5	6,25	2,52	0,000	0,01	0,86	
11,9	141,61	36,89	9,61	2,96	0,019	0,49	4,39	
14,1	198,81	46,53	10,89	3,35	0,003	0,81	1,60	
17,8	316,84	69,42	15,21	4,01	0,012	2,25	2,75	
69,7	**834,37**	**207,45**	**53,3**	**19,2**	**0,115**	**7,220**	**36,50**	
8,71	104,30	25,93	6,66	2,40	0,01		4,56	

Совершаем протягивание полученного результата по области «К4:К12»

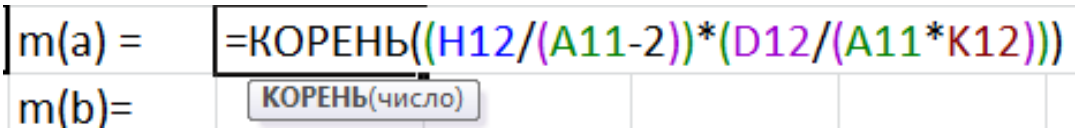
J	K

A	Q(x)
5,35	57,95
9,61	30,39
1,51	14,54
10,41	2,00
0,86	0,47
4,39	10,16
1,60	29,03
2,75	82,58
36,50	**227,11**

Теперь, имея все необходимые значения, найдем стандартные ошибки для *a* и *b*. В ячейку Н15 и Н16 введем соотвествующие комментарии «m_a=», «m_b=»

H
Q(e)
0,003
0,016
0,001
0,062
0,000
0,019
0,003
0,012
0,115
0,01
m(a) =
m(b)=

В ячейке «I15» записываем формулу:
«=КОРЕНЬ((H12/(A11-2)) *(D12/(A11*K12))) »



Нажав Enter, получим, что m_a = 0,094
Точно также вычислим m_b. В ячейку «I16» вписываем формулу «=КОРЕНЬ((H12/(A11-2))/K12)»

H	I	J	K
Q(e)	**Q**	**A**	**Q(x)**
0,003	1,96	5,35	57,95
0,016	1,21	9,61	30,39
0,001	0,49	1,51	14,54
0,062	0	10,41	2,00
0,000	0,01	0,86	0,47
0,019	0,49	4,39	10,16
0,003	0,81	1,60	29,03
0,012	2,25	2,75	82,58
0,115	**7,220**	**36,50**	**227,11**
0,01		4,56	
m(a) =	0,094		
m(b)=	=КОРЕНЬ((H12/(A11-2))/K12)		

Получим m_b = 0,0092

m(a) =	0,094
m(b)=	0,0092

Теперь, подставив в формулу найденные значения стандартных отклонений, вычислим статистические оценки коэффициентов регрессии.

В ячейке J15 введем комментарий «t_a = », далее в ячейку K15 введем формулу «=D15/I15»

C	D	E	F	G	H	I	J	K
X	**X^2**	**XY**	**Y^2**	**Y(t)**	**Q(e)**	**Q**	**A**	**Q(x)**
1,1	1,21	1,1	1	1,05	0,003	1,96	5,35	57,95
3,2	10,24	4,16	1,69	1,42	0,016	1,21	9,61	30,39
4,9	24,01	8,33	2,89	1,73	0,001	0,49	1,51	14,54
7,3	53,29	17,52	5,76	2,15	0,062	0	10,41	2,00
9,4	88,36	23,5	6,25	2,52	0,000	0,01	0,86	0,47
11,9	141,61	36,89	9,61	2,96	0,019	0,49	4,39	10,16
14,1	198,81	46,53	10,89	3,35	0,003	0,81	1,60	29,03
17,8	316,84	69,42	15,21	4,01	0,012	2,25	2,75	82,58
69,7	**834,37**	**207,45**	**53,3**	**19,2**	**0,115**	**7,220**	**36,50**	**227,11**
8,71	104,30	25,93	6,66	2,40	0,01		4,56	
a=	0,86		**R=**	**0,99**	m(a) =	0,094	t(a) =	=D15/I15
b=	0,18		R^2 =	0,98	m(b)=	0,0092		

Получим ta = 9,16 , точно также определим и статистику коэффициента регрессии b

a=	0,86	**R=**	**0,99**	m(a) =	0,094	t(a) =	9,15935
b=	0,18	R^2 =	0,98	m(b)=	0,0092	t(b) =	=D16/I16

t_b = 19,26

t(a) =	9,16
t(b) =	19,26

Сравниваем статистики соответствующих параметров с табличным значением t – критерия Стьюдента при уровне α = 0,05 и степени свободы (n – 2), равной 6. $t_{0,05(8-2)}$ = 2,45

$9{,}16 > 2{,}45 \rightarrow H_1$

$19{,}26 > 2{,}45 \rightarrow H_1$

Как видим оба параметра уравнения регрессии статистически значимы.

Точечный и интервальный прогноз

Предположим, нам необходимо определить объем денежной массы при уровне национального дохода, равном 18.

Для этого по условию точечного прогноза подставим в модель интересующее нас значение *x* (национальный доход):

$$\hat{y} = 0{,}86 + 0{,}18\, x_p = 0{,}86 + 0{,}18 * 18 = 4{,}04$$

Т.е. объем денежной массы при заданном уровне национального дохода составит 4,04 (млрд.).

m(a) =	0,094	**t(a) =**	9,16
m(b)=	0,0092	**t(b) =**	19,26
y(т.п.)=	=D14+D15*18		

В практической деятельности необходимо знать не только точечный прогноз среднего значения, а весь диапазон его возможных значений. Т.е. необходимо определить интервалы, в пределах которых значение данного прогноза может варьироваться. Определим пределы варьирования данного прогноза с 95% надежностью. Для этого воспользуемся формулой интервального прогноза:

$$\hat{y}_p - m_{\hat{y}_p} \cdot t_T \le \hat{y}_p \le \hat{y}_p + m_{\hat{y}_p} \cdot t_T$$

$$m_{\hat{y}_p} = \sqrt{\frac{\sum(y - \hat{y})^2}{n-2}} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{n \cdot \sum(x - \bar{x})^2}}$$

Рассчитаем дисперсию индивидуальных значений $m_{\hat{y}_p}$, иначе говоря, стандартную ошибку прогноза.

Все необходимые расчетные значения в нашей таблице уже вычислены, поэтому просто воспользуемся ими.

В ячейку H17 впишем комментарий $m_{\hat{y}_p}$, соответственно в ячейке I17 проведем все необходимые расчеты, впишем в нее формулу «=корень(H11/(A10-2))*корень(1+1/A10+(18-C12)^2/(A10*K11))»

m(y)=	=корень(H11/(A10-2))*корень(1+1/A10+(18-C12)^2/(A10*K11))

2	n	Y	X	X^2	XY	Y^2	Y(t)	Q(e)	Q	A	Q(x)
3	1	1	1,1	1,21	1,1	1	1,05	0,003	1,96	5,35	57,95
4	2	1,3	3,2	10,24	4,16	1,69	1,42	0,016	1,21	9,61	30,39
5	3	1,7	4,9	24,01	8,33	2,89	1,73	0,001	0,49	1,51	14,54
6	4	2,4	7,3	53,29	17,52	5,76	2,15	0,062	0	10,41	2,00
7	5	2,5	9,4	88,36	23,5	6,25	2,52	0,000	0,01	0,86	0,47
8	6	3,1	11,9	141,61	36,89	9,61	2,96	0,019	0,49	4,39	10,16
9	7	3,3	14,1	198,81	46,53	10,89	3,35	0,003	0,81	1,60	29,03
10	8	3,9	17,8	316,84	69,42	15,21	4,01	0,012	2,25	2,75	82,58
11	сумма	**19,2**	**69,7**	**834,37**	**207,45**	**53,3**	**19,2**	**0,115**	**7,220**	**36,50**	**227,11**
12	**среднее**	2,40	8,71	104,30	25,93	6,66	2,40	0,01		4,56	
13	дельта	1816,87									
14	дельта(a)	1560,64	a=	0,86		**R=**	**0,99**	**m(a) =**	0,094	**t(a) =**	9,16
15	дельта(b)	321,36	b=	0,18		R^2 =	0,98	**m(b)=**	0,0092	**t(b) =**	19,26
16						**F =**	**371,002**	y(т.п.)=	4,04273		
17						t(R) =	19,2614	**m(y)=**	0,15		

t– критерий Стьюдента при 95% значение надежности определен выше:

$t_{0,05(8-2)} = 2,45$. Впишем его в ячейку I18.

Теперь, имея все необходимые значения для расчета интервального прогноза, определим значения денежной массы $y_{(min)}$ и $y_{(max)}$ в ячейках J16 и J17, введя формулы «=I16+I18*I17» и «=I16-I18*I17».

4,04273	**Y(max)=**	=I16+I18*I17	
0,15			
2,45			

y(т.п.)=	4,04273	**Y(max)=**	4,40986
m(y)=	0,15	**Y(min)=**	=I16-I18*I17
t =	2,45		

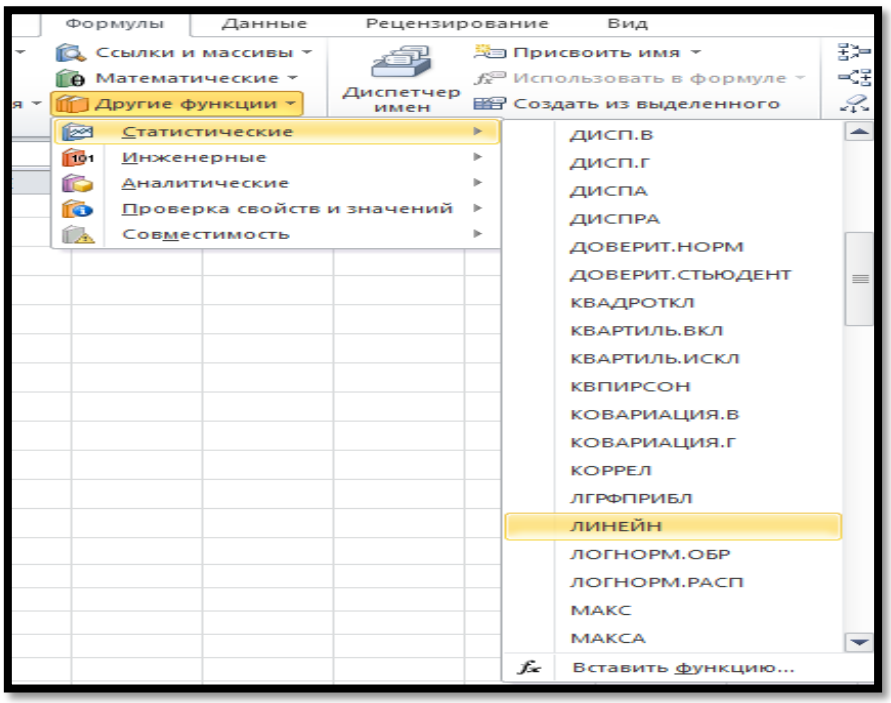
$$3,68 \leq \hat{y}_p \leq 4,41$$

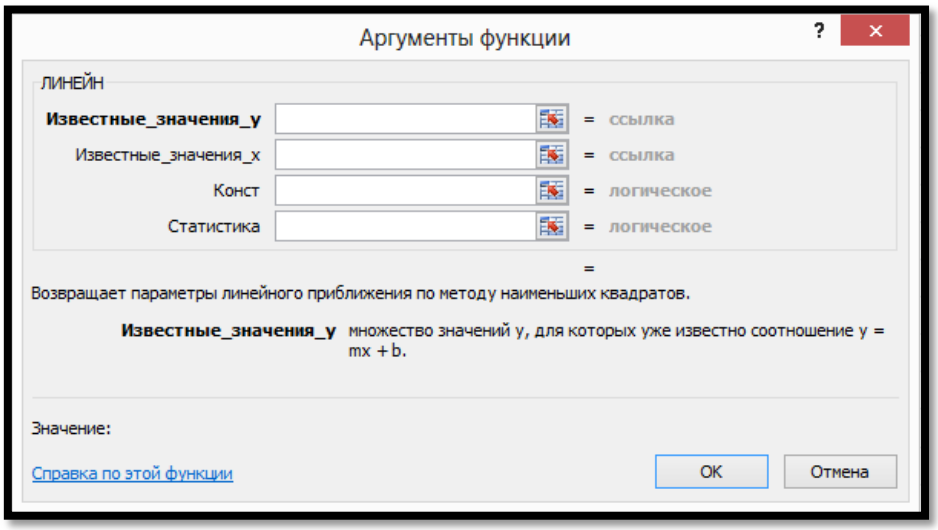
Т.е. объем денежной массы не опустится ниже отметки в 3,68(млрд.) и не поднимется выше 4,41(млрд.).

	A	B	C	D	E	F	G	H	I	J	K
1											
2	n	Y	X	X^2	XY	Y^2	Y(t)	Q(e)	Q	A	Q(x)
3	1	1	1,1	1,21	1,1	1	1,05	0,003	1,96	5,35	57,95
4	2	1,3	3,2	10,24	4,16	1,69	1,42	0,016	1,21	9,61	30,39
5	3	1,7	4,9	24,01	8,33	2,89	1,73	0,001	0,49	1,51	14,54
6	4	2,4	7,3	53,29	17,52	5,76	2,15	0,062	0	10,41	2,00
7	5	2,5	9,4	88,36	23,5	6,25	2,52	0,000	0,01	0,86	0,47
8	6	3,1	11,9	141,61	36,89	9,61	2,96	0,019	0,49	4,39	10,16
9	7	3,3	14,1	198,81	46,53	10,89	3,35	0,003	0,81	1,60	29,03
10	8	3,9	17,8	316,84	69,42	15,21	4,01	0,012	2,25	2,75	82,58
11	сумма	**19,2**	**69,7**	**834,37**	**207,45**	**53,3**	**19,2**	**0,115**	**7,220**	**36,50**	**227,11**
12	**среднее**	2,40	8,71	104,30	25,93	6,66	2,40	0,01		4,56	
13	дельта	1816,87									
14	дельта(a)	1560,64	a=	0,86		**R=**	**0,99**	**m(a) =**	0,094	**t(a) =**	9,16
15	дельта(b)	321,36	b=	0,18		R^2 =	0,98	**m(b)=**	0,0092	**t(b) =**	19,26
16						**F =**	**371,002**	**y(т.п.)=**	4,04273	**Y(max)=**	4,410
17						t(R) =	19,2614	**m(y)=**	0,15	**Y(min)=**	3,676
18								**t =**	2,45		

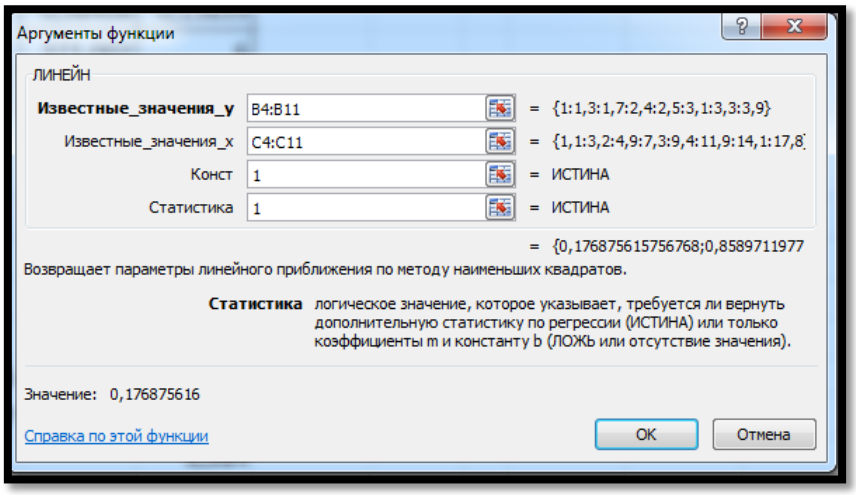
Встроенные средства MS Excel оценки параметров модели регрессии

Для быстрого решения задачи и нахождения параметров регрессионной модели выделим область в Excel 2×5 (B14:C18), далее откроем меню «Формулы» - «Статистические» - «Линейн»- «Ок»





В появившемся окне в окне ввода *y* выделяем ячейки «B4:B11», а в окне ввода значений *x*: «C4:C 11», в поле «константа» ставим 1, в поле «Статистика» 1(1 – истина, 0 – ложь).



Нажимаем – ок.

0,1769	

Далее нажать f**2** и комбинацию - **ctrl + shift + Enter**

0,1769	0,85897
0,0092	0,09378
0,9841	0,13839
371	6
7,1051	0,11491

В появившейся области решений первая строка является множеством коэффициентов регрессии, и соответственно модель примет вид:

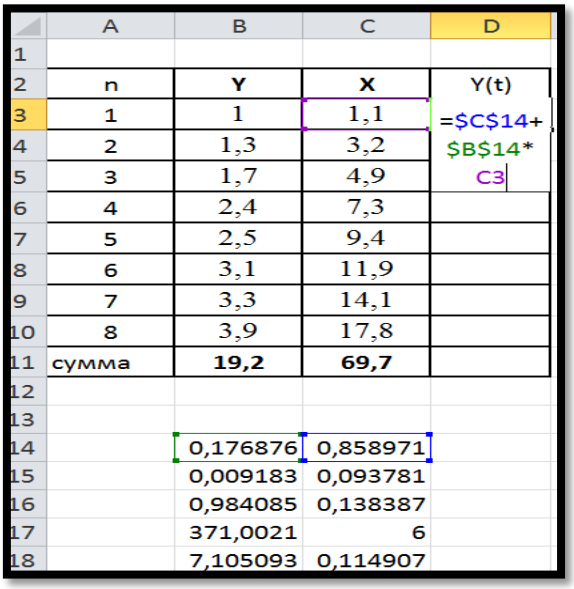
$$\hat{y} = 0,86 + 0,18x$$

Параметры построенной модели говорят о том, что с увеличением национального дохода на единицу измерения объем денежной массы увеличится на 0, 18 млрд. руб.

По результатам модели найдем значения наблюдений $\hat{y}$, подставив в модель значения *x*. Добавим для начала столбец *y(t)* в нашу таблицу.

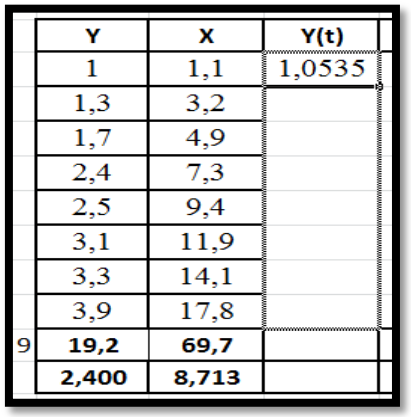
Y	X	Y(t)
1	1,1	
1,3	3,2	
1,7	4,9	
2,4	7,3	
2,5	9,4	
3,1	11,9	
3,3	14,1	
3,9	17,8	

Запишем в ячейку формулу для расчета регрессии.



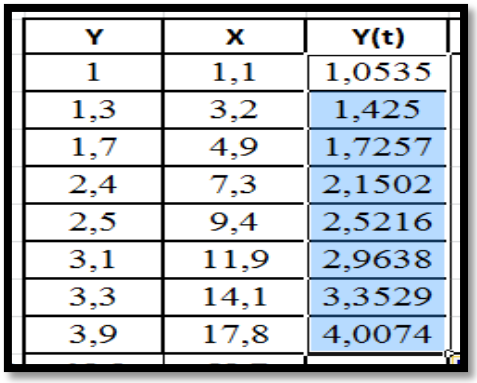
	A	B	C	D
1				
2	n	Y	X	Y(t)
3	1	1	1,1	=$C$14+
4	2	1,3	3,2	$B$14*
5	3	1,7	4,9	C3
6	4	2,4	7,3	
7	5	2,5	9,4	
8	6	3,1	11,9	
9	7	3,3	14,1	
10	8	3,9	17,8	
11	сумма	**19,2**	**69,7**	
12				
13				
14		0,176876	0,858971	
15		0,009183	0,093781	
16		0,984085	0,138387	
17		371,0021	6	
18		7,105093	0,114907	

Нажимаем Enter и совершаем протягивание для всего столбца.



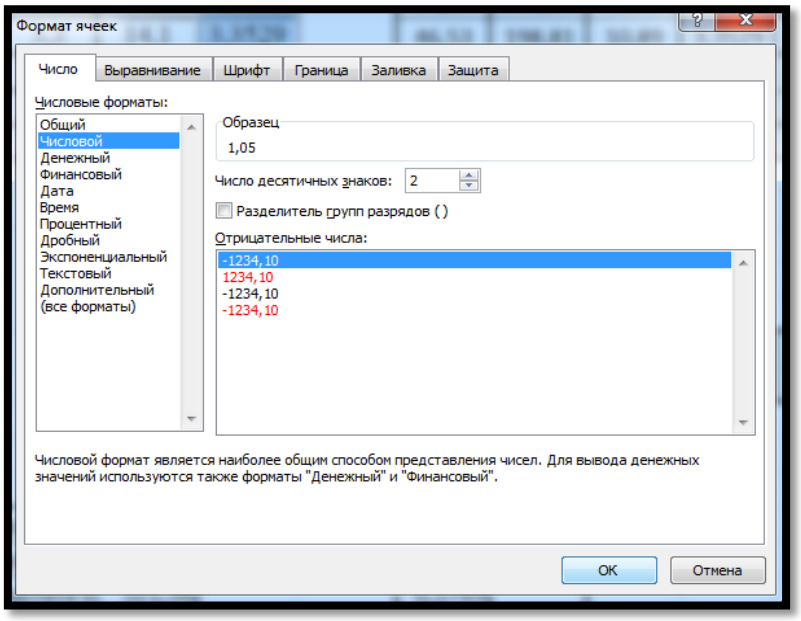
	Y	X	Y(t)
	1	1,1	1,0535
	1,3	3,2	
	1,7	4,9	
	2,4	7,3	
	2,5	9,4	
	3,1	11,9	
	3,3	14,1	
	3,9	17,8	
9	**19,2**	**69,7**	
	2,400	**8,713**	

Получим



Y	X	Y(t)
1	1,1	1,0535
1,3	3,2	1,425
1,7	4,9	1,7257
2,4	7,3	2,1502
2,5	9,4	2,5216
3,1	11,9	2,9638
3,3	14,1	3,3529
3,9	17,8	4,0074

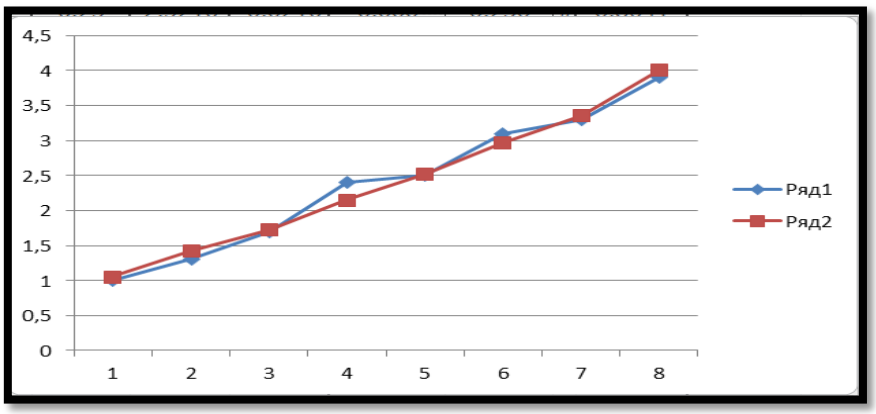
Округлим для простоты восприятия результаты в столбце *y(t)*, выделим столбец, вызовем контекстное меню – *формат ячеек – числовой*



Изобразим на диаграмме близость между наблюдениями *y* и *y(t)*

Y	X	Y(t)
1	1,1	1,05
1,3	3,2	1,42
1,7	4,9	1,73
2,4	7,3	2,15
2,5	9,4	2,52
3,1	11,9	2,96
3,3	14,1	3,35
3,9	17,8	4,01
19,2	**69,7**	**19,2**

Выделим эти столбцы, затем открываем меню «Вставка»-«Диаграммы» - «График»

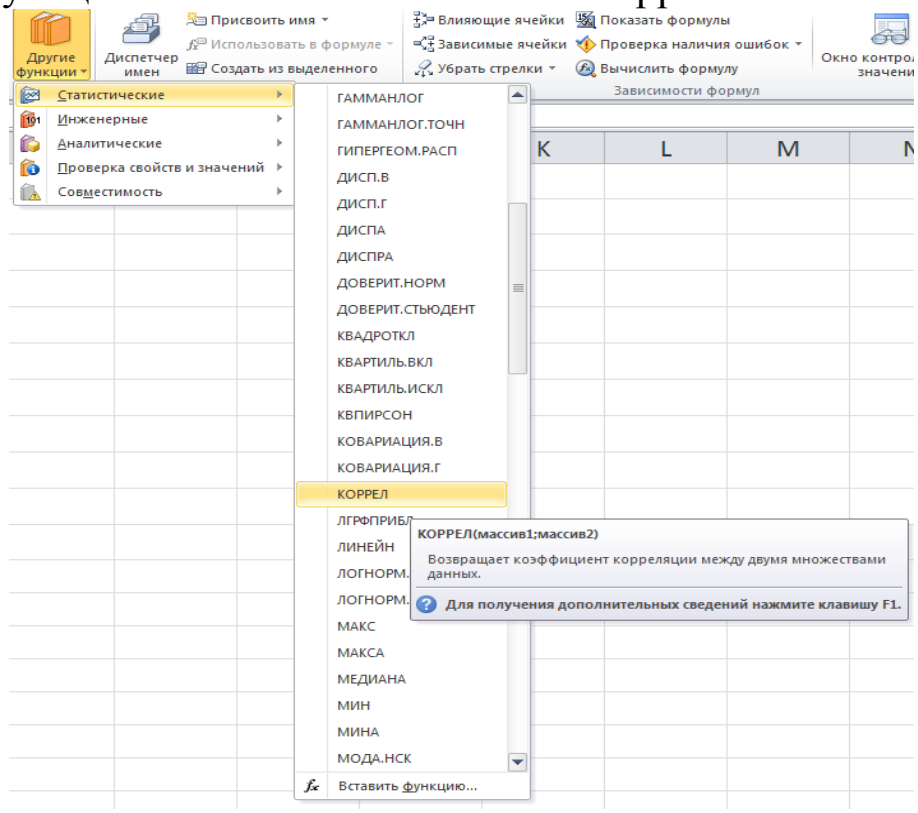


Как видим, графики рядов наблюдений фактических и теоретических наблюдений близки, в связи с чем можно предположить, что коэффициент корреляции будет давать значение близкое к единице.

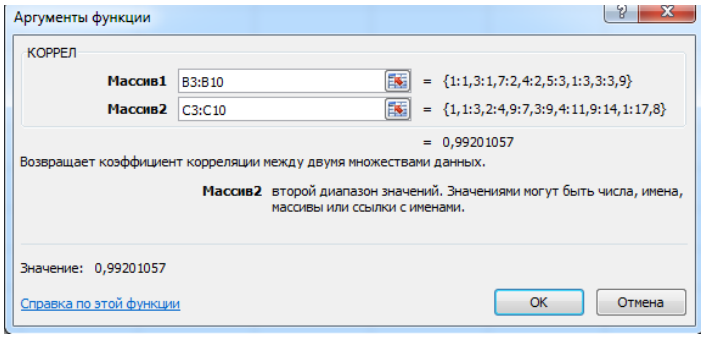
Найдем коэффициент корреляции.

Выделим ячейку F14, далее открываем меню – «формулы» -

«другие функции» - « статистические » - «коррел»



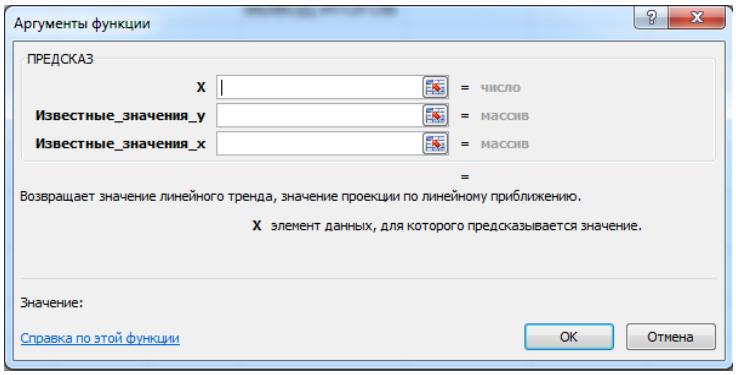
Заполняем открывшиеся поля:

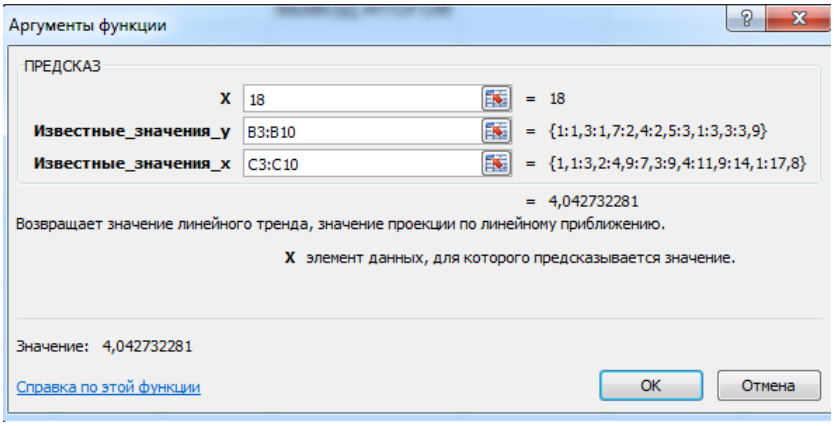


Нажимаем Ок , получим:

R=	0,992011

Получить для заданных наблюдений предсказанное значение *y* по известному *x*=18 можно функцией ПРЕДСКАЗ из категории «Статистические»





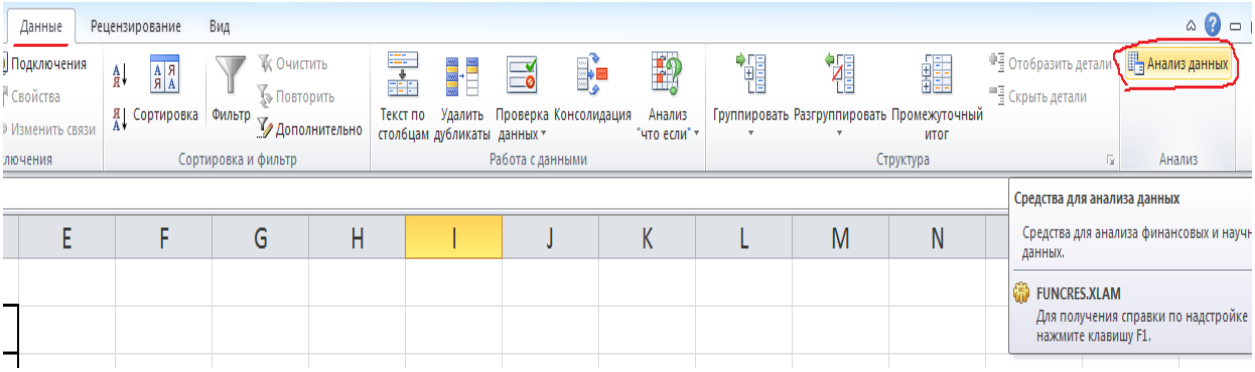
Покажем, что представляют собой остальные значения в области, которую мы выделили:

$\widehat{b}_1$	a
m_{b1}	m_a
R^2	$m_{\hat{y}}$
F	df
Q_r	Q_e

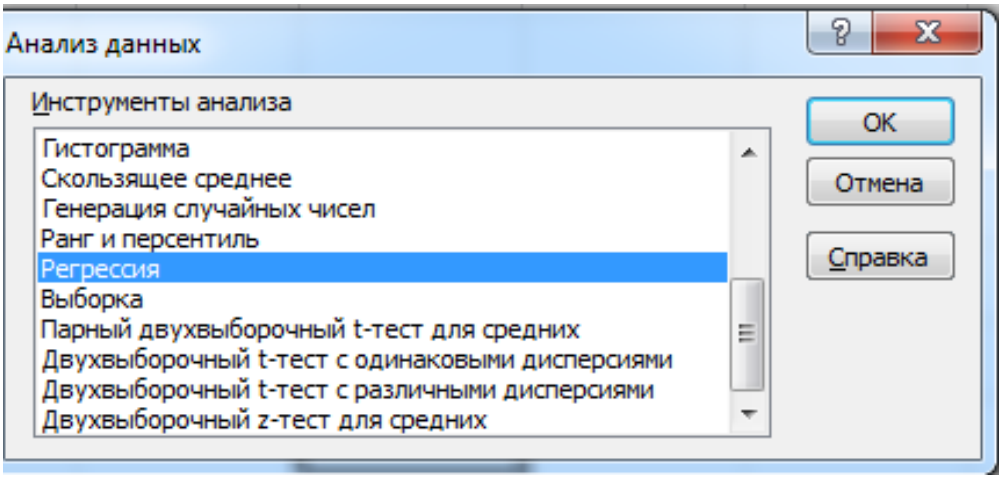
Как видим, необходимые вычисления при оценке значимости коэффициента корреляции, уравнения регрессии, коэффициентов регрессии, а так же стандартная ошибка прогноза при интервальном прогнозировании здесь уже имеются. Нам же остаётся лишь подставлять их в соответствующие формулы.

Еще одним инструментом при расчете регрессионной модели является - «Пакет анализа».

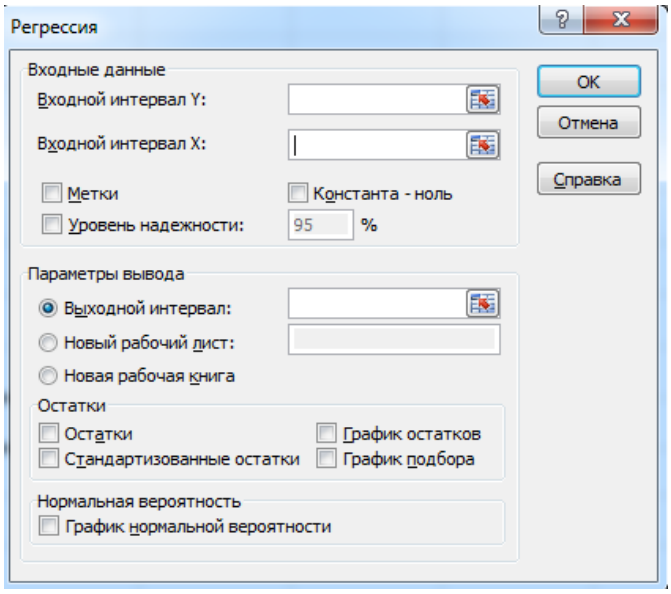
Открываем меню «Надстройки» - «пакет анализа» - «перейти» Далее в меню «Данные » появится «Анализ данных»



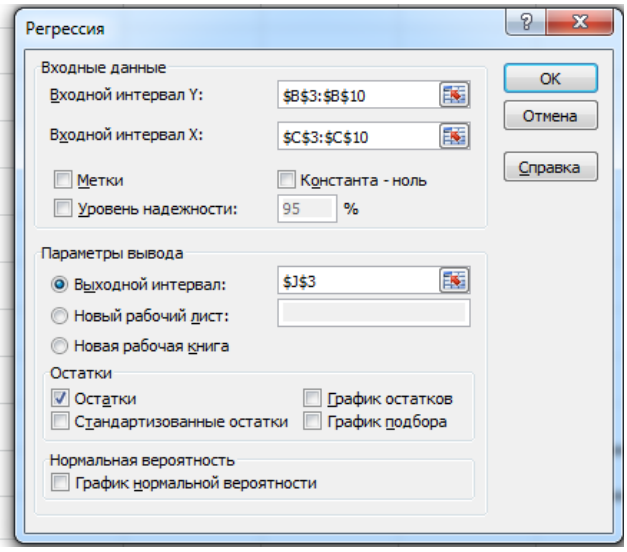
Запускаем его, открывается окно:



Заполняем поля в появившемся окне:



Y	X	Y(t)
1	1,1	1,05
1,3	3,2	1,42
1,7	4,9	1,73
2,4	7,3	2,15
2,5	9,4	2,52
3,1	11,9	2,96
3,3	14,1	3,35
3,9	17,8	4,01
19,2	**69,7**	**19,2**



Получим:

Регрессионная статистика	
Множественный R	0,9920
R-квадрат	0,9841
Нормированный R-квадрат	0,9814
Стандартная ошибка	0,1384
Наблюдения	8

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1	7,10509348	7,1050935	371,002122	0,000001
Остаток	6	0,11490652	0,0191511		
Итого	7	7,22			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	0,8589712	0,09378084	9,1593465	0,0001	0,629497751	1,08844
Переменная X 1	0,17687562	0,0091829	19,261415	0,00000	0,154405873	0,19935

Здесь приведены все наши расчеты, рассмотренные выше, с соответствующими разъяснениями, что является крайне большим преимуществом по сравнению с функцией «Линейн», так же здесь сразу найдены коэффициент корреляции и детерминации, рассчитаны статистические оценки *F*– Фишера и *t*– статистики Стьюдента для коэффициентов регрессии *a* и *b*.

Контрольные вопросы

1. Что понимается под регрессией в теории вероятностей и математической статистике?
2. Какие задачи решаются при построении уравнения регрессии?
3. Какие методы применяются для выбора вида модели регрессии?
4. Какие функции чаще всего используются для построения уравнения парной регрессии?
5. Какой вид имеет система нормальных уравнений метода наименьших квадратов?
6. Что при проверке статистических гипотез называют уровнем значимости?
7. Как проверяется значимость уравнения регрессии?
8. Как проверяется значимость коэффициентов уравнения регрессии?
9. Как вычисляется коэффициент детерминации R^2 ?
10. По какой формуле вычисляется выборочный коэффициент

парной корреляции r_{xy} ?

11. Как проверяется значимость выборочного коэффициента парной корреляции?

12. Как строится доверительный интервал для линейного коэффициента парной корреляции?

13. Как вычисляется и что показывает индекс детерминации?

14. Как осуществляется построение доверительного интервала прогноза в случае линейной регрессии?

15. Как вычисляется и как интерпретируется коэффициент эластичности Э?

Глава 3. Множественная регрессия и корреляция

Модель парной регрессии дает хорошие результаты при анализе и прогнозе в случае, когда влияние остальных факторов на объект исследования не существенно и ими можно пренебречь, не включив их в модель.

Однако, такие случаи редки и влиянием этих факторов по большей части пренебрегать нельзя. В этом случае их вводят в модель, дабы определить их влияние на объект исследования, т.е. необходимо построить уравнение множественной регрессии

$$y = \hat{f}(x_1, x_2, ..., x_m),$$

где y – объясняемая переменная (результативный признак),

x_i – объясняющие, или независимые, переменные (признаки-факторы).

Множественная регрессия получила широкое распространение при решении задач, связанных с исследованием проблем спроса, анализе доходности акций, при изучении функции издержек производства, в макроэкономических расчетах и в решении многих других вопросов эконометрики. На сегодняшний день множественная регрессия – один из наиболее популярных инструментов в эконометрике. Основной идеей, преследуемой множественной регрессией, является – построение модели с включенным в нее большим числом факторов, определяя их воздействие на моделируемый объект как по отдельности, так и в совокупности.

3.1. Спецификация модели. Отбор факторов при построении уравнения множественной регрессии

Первым этапом в построении уравнения множественной регрессии начинается с решения вопроса о спецификации модели.

На этапе спецификации модели учитывается два важных момента - отбор факторов и выбор вида уравнения регрессии.

Процесс включения какого – либо набора факторов в разрабатываемую модель множественной регрессии связан с мнением исследователя относительно природы взаимосвязи анализируемого показателя с остальными экономическими явлениями. Факторы, которые мы собираемся включить в модель должны отвечать следующим требованиям

1. Факторы должны иметь количественную характеристику. В

случае, когда нужно включить в модель качественный фактор, который не имеет количественное измерение, тогда данному фактору необходимо придать количественную определенность.

2. Факторы не должны иметь высокую корреляционную зависимость между собой, тем более находиться в точной функциональной связи.

Если включить в модель фактор с достаточно высокой интеркорреляцией, это может способствовать к появлению нежелательных последствий – система нормальных уравнений может оказаться плохо обусловленной и повлечь за собой неустойчивость и ненадежность оценок коэффициентов регрессии.

В случае наличия высокой корреляции между независимыми переменными нельзя будет выявить их изолированное влияние на результативный показатель и соответственно коэффициенты уравнения регрессии окажутся не интерпретируемыми.

Факторы, которые мы собираемся включать в модель, должны объяснить вариацию эндогенной переменной.

В случае, когда строят модель с m количеством факторов, тогда для нее вычисляется коэффициент детерминации R^2, который фиксирует долю объясненной вариации результативного признака за счет рассматриваемых в регрессии m факторов.

В случае дополнительного включения в модель регрессии $m+1$ факторов коэффициент детерминации должен возрастать:

$$R^2_{m+1} \geq R^2_m \ .$$

Ели данное условие не выполняется значит эти показатели не отличаются друг от друга, соответственно включенный в модель фактор x_{m+1} не только не улучшает модель, но и является лишним по сути фактором.

Процесс включения в модель ненужных факторов может привести к появлению статистической незначимости коэффициентов регрессии, определяемой при помощи критерия Стьюдента.

Таким образом, хотя теоретически регрессионная модель позволяет учесть любое число факторов, практически в этом нет необходимости. Отбор факторов производится на основе качественного теоретико-экономического анализа. Однако теоретический анализ часто не позволяет однозначно ответить на вопрос о количественной взаимосвязи рассматриваемых признаков и целесообразности включения фактора в модель. Поэтому, отбор факторов обычно осуществляется в две стадии: на первой

подбираются факторы исходя из сущности проблемы; на второй – на основе матрицы показателей корреляции определяют статистики для параметров регрессии.

Коэффициенты интеркорреляции (т.е. корреляции между объясняющими переменными) позволяют исключать из модели дублирующие факторы. Считается, что две переменные явно коллинеарны, т.е. находятся между собой в линейной зависимости, если $r_{x_i x_j} \geq 0,7$. Если факторы явно коллинеарны, то они дублируют друг друга и один из них рекомендуется исключить из регрессии. Предпочтение при этом отдается не фактору, более тесно связанному с результатом, а тому фактору, который при достаточно тесной связи с результатом имеет наименьшую тесноту связи с другими факторами. В этом требовании проявляется специфика множественной регрессии как метода исследования комплексного воздействия факторов в условиях их независимости друг от друга.

Пусть, например, при изучении зависимости $y = \hat{f}(x_1, x_2, x_3)$ матрица парных коэффициентов корреляции оказалась следующей:

Таблица 2.1

	y	x_1	x_2	x_3
y	1	0,8	0,7	0,6
x_1	0,8	1	0,8	0,5
x_2	0,7	0,8	1	0,2
x_3	0,6	0,5	0,2	1

Очевидно, что факторы x_1 и x_2 дублируют друг друга. В анализ целесообразно включить фактор x_2, а не x_1, хотя корреляция x_2 с результатом y слабее, чем корреляция фактора x_1 с y $\left(r_{yx_2} = 0,7 < r_{yx_1} = 0,8\right)$, но зато значительно слабее межфакторная корреляция $r_{x_2 x_3} = 0,2 < r_{x_1 x_3} = 0,5$. Поэтому в данном случае в уравнение множественной регрессии включаются факторы x_2, x_3.

По величине парных коэффициентов корреляции обнаруживается лишь явная коллинеарность факторов. Наибольшие трудности в использовании аппарата множественной регрессии возникают при наличии мультиколлинеарности факторов, когда более, чем два фактора связаны между собой линейной

зависимостью, т.е. имеет место совокупное воздействие факторов друг на друга. Наличие мультиколлинеарности факторов может означать, что некоторые факторы будут всегда действовать в унисон. В результате вариация в исходных данных перестает быть полностью независимой и нельзя оценить воздействие каждого фактора в отдельности.

Включение в модель мультиколлинеарных факторов нежелательно в силу следующих последствий:

1. Затрудняется интерпретация параметров множественной регрессии как характеристик действия факторов в «чистом» виде, ибо факторы коррелированы; параметры линейной регрессии теряют экономический смысл.

2. Оценки параметров ненадежны, обнаруживают большие стандартные ошибки и меняются с изменением объема наблюдений (не только по величине, но и по знаку), что делает модель непригодной для анализа и прогнозирования.

Для оценки мультиколлинеарности факторов может использоваться определитель матрицы парных коэффициентов корреляции между факторами.

Если бы факторы не коррелировали между собой, то матрица парных коэффициентов корреляции между факторами была бы единичной матрицей, поскольку все недиагональные элементы $r_{x_i x_j}$ $(i \neq j)$ были бы равны нулю. Так, для уравнения, включающего три объясняющих переменных

$$y = a + b_1 x_1 + b_2 x_2 + b_3 x_3,$$

матрица коэффициентов корреляции между факторами имела бы определитель, равный единице:

$$\operatorname{Det} \mathbf{R} = \begin{vmatrix} r_{x_1 x_1} & r_{x_1 x_2} & r_{x_1 x_3} \\ r_{x_2 x_1} & r_{x_2 x_2} & r_{x_2 x_3} \\ r_{x_3 x_1} & r_{x_3 x_2} & r_{x_3 x_3} \end{vmatrix} = \begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix} = 1.$$

Если же, наоборот, между факторами существует полная линейная зависимость и все коэффициенты корреляции равны единице, то определитель такой матрицы равен нулю:

$$\mathrm{Det}\,\mathbf{R} = \begin{vmatrix} r_{x_1x_1} & r_{x_1x_2} & r_{x_1x_3} \\ r_{x_2x_1} & r_{x_2x_2} & r_{x_2x_3} \\ r_{x_3x_1} & r_{x_3x_2} & r_{x_3x_3} \end{vmatrix} = \begin{vmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{vmatrix} = 0.$$

Чем ближе к нулю определитель матрицы межфакторной корреляции, тем сильнее мультиколлинеарность факторов и ненадежнее результаты множественной регрессии. И, наоборот, чем ближе к единице определитель матрицы межфакторной корреляции, тем меньше мультиколлинеарность факторов.

Существует ряд подходов преодоления сильной межфакторной корреляции. Самый простой путь устранения мультиколлинеарности состоит в исключении из модели одного или нескольких факторов. Другой подход связан с преобразованием факторов, при котором уменьшается корреляция между ними.

Одним из путей учета внутренней корреляции факторов является переход к совмещенным уравнениям регрессии, т.е. к уравнениям, которые отражают не только влияние факторов, но и их взаимодействие. Так, если $y = f(x_1, x_2, x_3)$, то возможно построение следующего совмещенного уравнения:

$$y = a + b_1x_1 + b_2x_2 + b_3x_3 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{23}x_2x_3 + \varepsilon.$$

Рассматриваемое уравнение включает взаимодействие первого порядка (взаимодействие двух факторов). Возможно включение в модель и взаимодействий более высокого порядка, если будет доказана их статистическая значимость по F-критерию Фишера, но, как правило, взаимодействия третьего и более высоких порядков оказываются статистически незначимыми.

Отбор факторов, включаемых в регрессию, является одним из важнейших этапов практического использования методов регрессии. Подходы к отбору факторов на основе показателей корреляции могут быть разные. Они приводят построение уравнения множественной регрессии соответственно к разным методикам. В зависимости от того, какая методика построения уравнения регрессии принята, меняется алгоритм ее решения на ЭВМ.

Наиболее широкое применение получили следующие методы построения уравнения множественной регрессии:

1. Метод исключения – отсев факторов из полного его набора.
2. Метод включения – дополнительное введение фактора.

3. Шаговый регрессионный анализ – исключение ранее введенного фактора.

При отборе факторов также рекомендуется пользоваться следующим правилом: число включаемых факторов обычно в 6–7 раз меньше объема совокупности, по которой строится регрессия. Если это соотношение нарушено, то число степеней свободы остаточной дисперсии очень мало. Это приводит к тому, что параметры уравнения регрессии оказываются статистически незначимыми, а F-критерий меньше табличного значения.

3.2. Метод наименьших квадратов (МНК). Свойства оценок на основе МНК

Множественная регрессия может иметь как линейный, так и нелинейный характер связи между факторами.

Линейный вид модели наиболее популярен ввиду наличия четкой интерпретации ее параметров.

В линейной множественной регрессии $y_x = a + b_1x_1 + b_2x_2 + ... + b_mx_m$ параметры при x называются коэффициентами «чистой» регрессии.

Параметры при x характеризуют среднее изменение y при изменении соответствующего фактора x на единицу своего измерения при условии, что значение других факторов неизменно.

Рассмотрим линейную модель множественной регрессии

$$y = a + b_1x_1 + b_2x_2 + ... + b_mx_m + \varepsilon. \qquad (2.1)$$

Классический подход к оцениванию параметров линейной модели множественной регрессии основан на методе наименьших квадратов (МНК). МНК позволяет получить такие оценки параметров, при которых сумма квадратов отклонений фактических значений результативного признака y от расчетных y минимальна:

$$\sum_i \left(y_i - y_{x_i} \right)^2 \rightarrow \min. \qquad (2.2)$$

Мы знаем, что по второй теореме Вейерштрасса функция достигает своего максимума или минимума в точке, в которой её производная равна нулю, значит, необходимо найти частные производные по каждому из параметров регрессии и приравнять их к нулю.

Итак, имеем функцию $m+1$ аргумента:

$$S(a,\ b_1,\ b_2,\ ...,\ b_m) = \sum \left(y - a - b_1x_1 - b_2x_2 - ... - b_mx_m \right)^2.$$

Находим частные производные первого порядка:

$$\begin{cases} \dfrac{\partial S}{\partial a} = -2\Sigma(y-a-b_1 x_1-b_2 x_2-\ldots-b_m x_m)=0; \\ \dfrac{\partial S}{\partial b_1} = -2\Sigma(y-a-b_1 x_1-b_2 x_2-\ldots-b_m x_m)\,x_1=0; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \dfrac{\partial S}{\partial b_m} = -2\Sigma(y-a-b_1 x_1-b_2 x_2-\ldots-b_m x_m)\,x_m=0. \end{cases}$$

После элементарных преобразований приходим к системе линейных нормальных уравнений для нахождения параметров линейного уравнения множественной регрессии (2.1):

$$\begin{cases} na+b_1\sum x_1+b_2\sum x_2+\ldots+b_m\sum x_m=\sum y, \\ a\sum x_1+b_1\sum x_1^2+b_2\sum x_1x_2+\ldots+b_m\sum x_1x_m=\sum yx_1, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a\sum x_m+b_1\sum x_1x_m+b_2\sum x_2x_m+\ldots+b_m\sum x_m^2=\sum yx_m. \end{cases}$$

Для двухфакторной модели данная система будет иметь вид:

$$\begin{cases} na+b_1\sum x_1+b_2\sum x_2=\sum y, \\ a\sum x_1+b_1\sum x_1^2+b_2\sum x_1x_2=\sum yx_1, \\ a\sum x_2+b_1\sum x_1x_2+b_2\sum x_2^2=\sum yx_2. \end{cases}$$

Оценка параметров множественной регрессии в матричном виде.

В матричном виде уравнение множественной регрессии выглядит следующим образом:

$Y = \beta \cdot X + \varepsilon,$

где У – матрица наблюдений *y*;

β – матрица всех параметров уравнения регрессии

X - матрица всех наблюдений переменной x_i.

$$\beta = (X^T X)^{-1} \cdot X^T Y$$

$(X^T X)^{-1}$ - обратная матрица

Обозначим $(X^T X)^{-1}$ через A^{-1} , тогда $\beta = A^{-1} \cdot X^T Y$

$A^{-1} = \frac{\bar{A}}{|A|}$; $\beta = \frac{\bar{A} \cdot X^T Y}{|A|}$,

где $\bar{A}$ - присоединенная матрица, $|A| -$ определитель матрицы $X^T X$

$$X^T X = \begin{vmatrix} n & \sum x_1 & \sum x_2 & \dots & \sum x_n \\ \sum x_1 & \sum x_1^2 & \sum x_1 x_2 & \dots & \sum x_1 x_n \\ \sum x_2 & \sum x_1 x_2 & \sum x_2^2 & \dots & \sum x_2 x_n \\ \dots & \dots & \dots & \dots & \dots \\ \sum x_n & \sum x_1 x_n & \sum x_2 x_n & \dots & \sum x_n^2 \end{vmatrix} \quad X^T Y = \begin{vmatrix} \sum y \\ \sum y x_1 \\ \sum y x_2 \\ \dots \\ \sum y x_n \end{vmatrix}$$

Элементы присоединенной матрицы представляют из себя алгебраические дополнения $a_{ij} = (-1)^{i+j} \cdot |M_{ij}|$, где M_{ij} - минор матрицы.

Пример:

Пусть объем выпуска некоторой продукции Y предприятия имеет линейную зависимость от затрат на рекламу X_1(тыс. руб.) и зарплаты X2 (тыс. руб.) сотрудников данного предприятия, занятых в производстве данной продукции. Вычислим коэффициенты уравнения линейной регрессии.

Y	**X_1**	**X_2**
21	11	13
34	14	11
29	19	10
44	24	9
59	38	8
68	35	8
74	42	7
88	33	5
103	36	5
109	53	6

$$X^T X = \begin{vmatrix} n & \sum x_1 & \sum x_2 \\ \sum x_1 & \sum x_1^2 & \sum x_1 x_2 \\ \sum x_2 & \sum x_1 x_2 & \sum x_2^2 \end{vmatrix} \quad X^T Y = \begin{vmatrix} \sum Y \\ \sum Y X_1 \\ \sum Y X_2 \end{vmatrix}$$

Дополним таблицу:

Y	**X_1**	**X_2**	**X_1X_2**	**X_1^2**	**X_2^2**	**YX_1**	**YX_2**
21	11	13	143	121	169	231	273
34	14	11	154	196	121	476	374
29	19	10	190	361	100	551	290
44	24	9	216	576	81	1056	396
59	38	8	304	1444	64	2242	472
68	35	8	280	1225	64	2380	544
74	42	7	294	1764	49	3108	518
88	33	5	165	1089	25	2904	440
103	36	5	180	1296	25	3708	515
109	53	6	318	2809	36	5777	654
629	**305**	**82**	**2244**	**10881**	**734**	**22433**	**4476**

Заполнив таблицу, получим матрицу:

$$X^T X = \begin{vmatrix} 10 & 305 & 82 \\ 305 & 10881 & 2244 \\ 82 & 2244 & 734 \end{vmatrix},$$

определитель которой равен:

$$A = \begin{vmatrix} 10 & 305 & 82 \\ 305 & 10881 & 2244 \\ 82 & 2244 & 734 \end{vmatrix} = 311866$$

Найдем элементы присоединенной матрицы:

$$a_{11} = (-1)^{1+1} \cdot \begin{vmatrix} 10881 & 2244 \\ 2244 & 734 \end{vmatrix} = 295118$$

$$a_{12} = -39862, a_{13} = -207822, a_{21} = -39862, a_{22} = 616, a_{23} = 2570$$

$$a_{31} = -207822, a_{32} = 2570, a_{33} = 15785.$$

Разделив элементы присоединенной матрицы на определитель, получим обратную матрицу:

$$A^{-1} = \frac{\bar{A}}{|A|} = \begin{vmatrix} 9{,}46278 & -0{,}1278 & -0{,}6664 \\ -0{,}1278 & 0{,}00198 & 0{,}0082 \\ -0{,}6664 & 0{,}00824 & 0{,}0506 \end{vmatrix}$$

$$X^T Y = \begin{vmatrix} 629 \\ 22433 \\ 4476 \end{vmatrix}$$

Перемножив две матрицы $A^{-1} \cdot X^T Y$, получим матрицу коэффициентов уравнения регрессии:

$$\beta = (X^T X)^{-1} \cdot X^T Y = \begin{vmatrix} 102.024 \\ 0{,}798 \\ -7{,}739 \end{vmatrix}$$

Отсюда уравнение регрессии имеет вид:

$$\hat{y} = 102{,}024 + 0{,}798x_1 - 7{,}739x_2$$

Данное уравнение показывает нам, что в случае увеличении одних лишь затрат на рекламу x_1 (при закреплении x_2 на неизменном уровне) на 1 тыс. руб. предложение выпускаемой продукции y увеличится в среднем на 0,789 т., а при увеличении затрат на оплату труда работников x_2 (при неизменном x_1) на 1 тыс. руб. объем выпуска продукции уменьшится в среднем на 7,739 т.

Множественная корреляция

Практическая значимость уравнения множественной регрессии оценивается с помощью показателя множественной корреляции и его квадрата — коэффициента детерминации.

Значение множественного коэффициента корреляции отражает тесноту связи оцениваемого набора факторов с исследуемым признаком, или же характеризует тесноту группового влияния рассматриваемого набора факторов на результат.

Показатель множественной корреляции может быть определен как индекс множественной корреляции по следующей формуле:

$$R_{yx1x2\ldots xp} = \sqrt{1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}}$$

Индекс множественной корреляции меняется в пределах от 0 до 1. Связь между результативным признаком (y) и набором факторов x будет тесной в случае, когда индекс множественной корреляции принимает значения, близкие к 1. И наоборот, чем ближе значение индекс множественной корреляции к 0, тем связь между результативным признаком (y) и набором факторов x слабее.

Пример: Рассчитаем множественный коэффициент корреляции для примера, рассмотренного выше.

Y	X_1	X_2	$\widehat{Y}$	$(Y-\widehat{Y})^2$	$(Y-\overline{Y})^2$
21	11	13	10,19	116,80	1755,61
34	14	11	28,06	35,23	835,21
29	19	10	39,79	116,50	1149,21
44	24	9	51,52	56,58	357,21
59	38	8	70,43	130,70	15,21
68	35	8	68,04	0,00	26,01
74	42	7	81,36	54,22	123,21
88	33	5	89,66	2,76	630,01
103	36	5	92,05	119,82	1608,01
109	53	6	97,88	123,66	2125,21
629	**305**	**82**	**629**	**756,264**	**8624,9**

$$R = \sqrt{1 - \frac{(\boldsymbol{Y} - \widehat{\boldsymbol{Y}})^{\boldsymbol{2}}}{(\boldsymbol{Y} - \overline{\boldsymbol{Y}})^{\boldsymbol{2}}}} = \sqrt{1 - \frac{756{,}264}{8624{,}9}} = 0{,}955$$

Делаем вывод о сильной связи между Y и совокупным влиянием факторов X_i.

Проверка существенности факторов и показатели качества регрессии

Коэффициент детерминации характеризует качество построенной модели в целом. Вычисляется как квадрат индекса множественной корреляции. Он характеризует долю дисперсии результативного признака, объясненную построенной моделью регрессии, с учетом влияния всего набора факторов, включенных в модель, выраженную в процентах.

Определяется, как и в случае с парной регрессией, по следующей формуле:

$$R^2 = 1 - \frac{(\boldsymbol{Y} - \widehat{\boldsymbol{Y}})^{\boldsymbol{2}}}{(\boldsymbol{Y} - \overline{\boldsymbol{Y}})^{\boldsymbol{2}}}$$

Чем ближе значение детерминации к единице, тем качественнее построена модель.

Коэффициент детерминации, найденный на основе предыдущего ряда данных, будет равен $R^2 = 0.91$. Ввиду того, что коэффициент принимает значение близкое к единице предполагаем, что полученная модель качественно описывает зависимость предложения продукции (Y) от затрат на рекламу (X_1) и зарплаты работников (X_2).

3.3. Значимость множественной регрессии и ее коэффициентов

Оценка значимости уравнения регрессии

Значимость уравнения множественной регрессии в целом, так же как и в парной регрессии, оценивается с помощью *F*-критерия Фишера:

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

$$\begin{cases} F_{\text{кр}} > F_T \rightarrow H_1 \\ F_{\text{кр}} < F_T \rightarrow H_0 \end{cases}$$

Оценим значимость уравнения множественной регрессии для рассматриваемого примера:

$$F = \frac{0.91}{1 - 0.91} \cdot \frac{10 - 2 - 1}{2} = 36.42$$

Значение *F*-критерия Фишера превосходит табличный показатель F_T-критерия =5,32

$F_{\text{кр}} > F_T \rightarrow H_1$

Принимается гипотеза H_1, согласно которой делаем вывод о значимости уравнения регрессии.

Оценка значимости коэффициентов множественной регрессии.

Оценка статистической значимости коэффициентов регрессии проводится по *t*–критерию Стьюдента.

Вычисляются *t* –статистики по каждому из показателей уравнения регрессии (a, b_1,b_2, …,b_n).

$$t_a = \frac{a}{m_a}, t_{b1} = \frac{b_1}{m_{b1}}, t_{b2} = \frac{b_2}{m_{b2}}, \dots, t_{bn} = \frac{b_n}{m_{bn}}.$$

S_{bj}- матрица всех стандартных ошибок коэффициентов множественной регрессии.

$$S_{bj} = \sqrt{S_e^2 \cdot |X^T X|^{-1}}$$

$$S_e^2 = \frac{(y - \hat{y})^2}{n - m - 1}$$

Из элементов обратной матрицы выбираем элементы, находящиеся по диагонали.

Оценим статистическую значимость коэффициентов регрессии:

$$S_e^2 = \frac{(y - \hat{y})^2}{n - m - 1} = \frac{756{,}264}{10 - 2 - 1} = 108{,}038$$

$$S_{bj} = \sqrt{S_e^2 \cdot |X^T X|^{-1}} = \sqrt{108{,}038 \cdot \begin{vmatrix} 9{,}463 \\ 0{,}002 \\ 0{,}051 \end{vmatrix}} = \begin{vmatrix} 31{,}974 \\ 0{,}461 \\ 2{,}338 \end{vmatrix}$$

Подставим найденные значения

$$t_a = \frac{a}{m_a} = \frac{102{,}024}{31{,}974} = 3{,}191, t_{b1} = \frac{b_1}{m_{b1}} = \frac{0{,}798}{0{,}461} = 1{,}727,$$

$$t_{b2} = \frac{b_2}{m_{b2}} = \frac{7{,}739}{2{,}338} = 3{,}31$$

Табличное значение t – критерия Стьюдента при уровне значимости, равном 0,05, и степени свободы V =7 (n – m – 1) составит t_T = 2,36.

Приходим к выводу, что коэффициент регрессии b_2 значим, а коэффициент b_1, напротив, оказался незначимым, ввиду того, что *t*–статистика данного параметра меньше табличного значения *t*-распределения Стьюдента.

3.4. Регрессионные модели с фиктивной переменной

В процессе изучения моделей множественной регрессии, мы предположили, что независимые переменные имеют числовой эквивалент. Однако в некоторых случаях в нашу модель приходится включить категориальную переменную. Помимо данных переменных, которые имеют чисто числовой характер, есть смысл попытаться учесть в модели такой фактор, который нельзя отразить числом, однако его присутствие и роль в самом процессе значительна. К примеру рассмотрим фактор, характеризующий расположение товара внутри магазина (к примеру, на витрине или нет). С целью учета в модели регрессии категориальную переменную стоит включить в нее фиктивную переменную. К примеру, если некая категориальная переменная будет иметь 2 категории, для ее представления хватает 1 фиктивной переменной X_d: $X_d = 0$, если наблюдение принадлежит 1-й категории, $X_d = 1$, если наблюдение принадлежит 2-й категории.

В качестве иллюстрации включения фиктивной переменной рассмотрим модель, в которой требуется предсказать среднюю оценочную стоимость недвижимости при помощи информации о 15 домах. В качестве независимых факторов возьмем размер жилой площади, в качестве фиктивной переменной включим наличие в доме камина X_2. X_2 будет равно нулю если камина нет, X_2 равно единице, если в доме камин имеется.

Цена, тыс. дол.	Размер, тыс. кв. фунтов.	Наличие камина
84,4	2	1
77,4	1,71	0
75,7	1,45	0
85,9	1,76	1
79,1	1,93	0
70,4	1,2	1
75,8	1,55	1
85,9	1,93	1
78,5	1,59	1
79,2	1,5	1
86,7	1,9	1
79,3	1,39	1
74,5	1,54	0
83,8	1,89	1
76,8	1,59	0

Сделаем предположение относительно того, что наклон оценочной стоимости, имеющий зависимость от площади дома, будет одинаковым как у домов, имеющих камин, так и не имеющих. Отсюда многофакторная модель регрессии имеет вид:

$$\widehat{\boldsymbol{Y}} = \boldsymbol{a} + \boldsymbol{b_1 x_1} + \boldsymbol{b_2 x_2} + \boldsymbol{\varepsilon},$$

где Y_i — прогнозируемая стоимость i-го дома, x_1 – площадь дома, b_1 - среднее увеличение стоимости дома от площади x_1, b_2 – коэффициент прироста оценочной стоимости дома при наличии в нем камина при постоянной величине площади дома.

Y	X_1	X_2	YX_1	YX_2	X_1X_2	X_1^2	X_2^2
84,4	2	1	168,8	84,4	2	4,00	1
77,4	1,71	0	132,35	0	0	2,92	0
75,7	1,45	0	109,77	0	0	2,10	0
85,9	1,76	1	151,18	85,9	1,76	3,10	1
79,1	1,93	0	152,66	0	0	3,72	0
70,4	1,2	1	84,48	70,4	1,2	1,44	1
75,8	1,55	1	117,49	75,8	1,55	2,40	1
85,9	1,93	1	165,79	85,9	1,93	3,72	1
78,5	1,59	1	124,82	78,5	1,59	2,53	1
79,2	1,5	1	118,8	79,2	1,5	2,25	1
86,7	1,9	1	164,73	86,7	1,9	3,61	1
79,3	1,39	1	110,23	79,3	1,39	1,93	1
74,5	1,54	0	114,73	0	0	2,37	0
83,8	1,89	1	158,38	83,8	1,89	3,57	1
76,8	1,59	0	122,11	0	0	2,53	0
1193,4	**24,93**	**10**	**1996,3**	**809,9**	**16,71**	**42,21**	**10**

Решение идентично рассмотренному выше:

$$X^T X = \begin{vmatrix} 15 & 24{,}93 & 10 \\ 24{,}93 & 42{,}21 & 16{,}71 \\ 10 & 16{,}71 & 10 \end{vmatrix}$$

$$|A| = 38{,}62$$

$$A^{-1} = \frac{\bar{A}}{|A|} = \begin{vmatrix} 3{,}70 & -2{,}13 & -0{,}14 \\ -2{,}13 & 1{,}29 & -0{,}03 \\ -0{,}14 & -0{,}03 & 0{,}30 \end{vmatrix}$$

$$X^T Y = \begin{vmatrix} 1193{,}4 \\ 1996{,}3 \\ 809{,}9 \end{vmatrix}$$

$$\beta = (X^T X)^{-1} \cdot X^T Y = \begin{vmatrix} 3{,}70 & -2{,}13 & -0{,}14 \\ -2{,}13 & 1{,}29 & -0{,}03 \\ -0{,}14 & -0{,}03 & 0{,}30 \end{vmatrix} \cdot \begin{vmatrix} 1193{,}4 \\ 1996{,}3 \\ 809{,}9 \end{vmatrix} = \begin{vmatrix} 50{,}09 \\ 16{,}19 \\ 3{,}85 \end{vmatrix}$$

$$\hat{У} = \mathbf{50{,}09 + 16{,}186X_1 + 3{,}853X_2},$$

где X_2 принимает лишь значения, равные нулю и единице.

В данной регрессионной модели коэффициенты регрессии интерпретируются следующим образом:

1. В случае, если переменная $\mathbf{X_2}$ неизменна, увеличение площади дома на 1000 кв. м. приведет к росту предсказанной оценочной стоимости в среднем на 16,2 тыс. долл.
2. В случае, если площадь дома неизменна, тогда наличие камина в доме будет способствовать росту его оценочной стоимости в среднем на 3,85 тыс. долл.

Найдем значение коэффициента корреляции:

Y	$\mathbf{X_1}$	$\mathbf{X_2}$	$\hat{\boldsymbol{Y}}$	$(Y - \hat{Y})^2$	$(Y - \bar{Y})^2$
84,4	2	1	86,32	3,67	23,4256
77,4	1,71	0	77,77	0,14	4,6656
75,7	1,45	0	73,56	4,58	14,8996
85,9	1,76	1	82,43	12,04	40,1956
79,1	1,93	0	81,33	4,97	0,2116
70,4	1,2	1	73,37	8,80	83,9056
75,8	1,55	1	79,03	10,44	14,1376
85,9	1,93	1	85,18	0,52	40,1956
78,5	1,59	1	79,68	1,39	1,1236
79,2	1,5	1	78,22	0,96	0,1296
86,7	1,9	1	84,70	4,01	50,9796
79,3	1,39	1	76,44	8,17	0,0676
74,5	1,54	0	75,02	0,27	25,6036
83,8	1,89	1	84,53	0,54	17,9776

76,8	1,59	0	75,83	0,95	7,6176
1193,4	**24,93**	**10**	**1193,4**	**61,432**	**325,14**

$$R = \sqrt{1 - \frac{(\boldsymbol{Y} - \widehat{\boldsymbol{Y}})^{\boldsymbol{2}}}{(\boldsymbol{Y} - \overline{\boldsymbol{Y}})^{\boldsymbol{2}}}} = \sqrt{1 - \frac{61{,}432}{325{,}14}} = 0{,}90$$

Коэффициент множественной корреляции показывает, что имеется тесная связь между ценой квартиры (Y) и совокупным влиянием на нее факторов (X_i).

Так же стоит обратить внимание на R^2, который равен – 0,81, что говорит о том, что стоимость квартиры на 81% зависит от площади и наличия камина.

Статистическая оценка значимости коэффициентов регрессии дает положительный результат, ибо *t*–статистика обоих коэффициентов превышает его табличное значение. Соответственно, каждая из рассмотренных переменных вносит значимый вклад в модель регрессии.

3.5.Средства MS Excel в множественном регрессионном анализе

Рассмотрим задачу моделирования годового товарооборота (Y) в зависимости от торговой площади (X_1) и среднего числа посетителей в день(X_2):

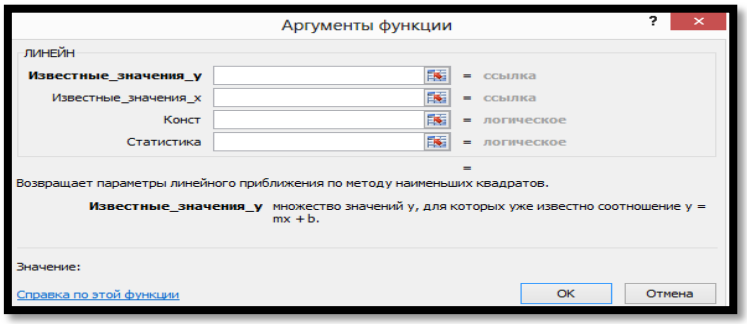
y	**x_1**	**x_2**
19,76	0,24	8,25
38,09	0,31	10,24
40,95	0,55	9,31
41,08	0,48	11,01
56,29	0,78	8,54
68,51	0,98	7,51
75,01	0,94	12,36
89,05	1,21	10,81
91,13	1,29	9,89
91,26	1,12	13,72
99,84	1,29	12,27
108,55	1,49	13,92

Введем данные в табличный процессор Excel

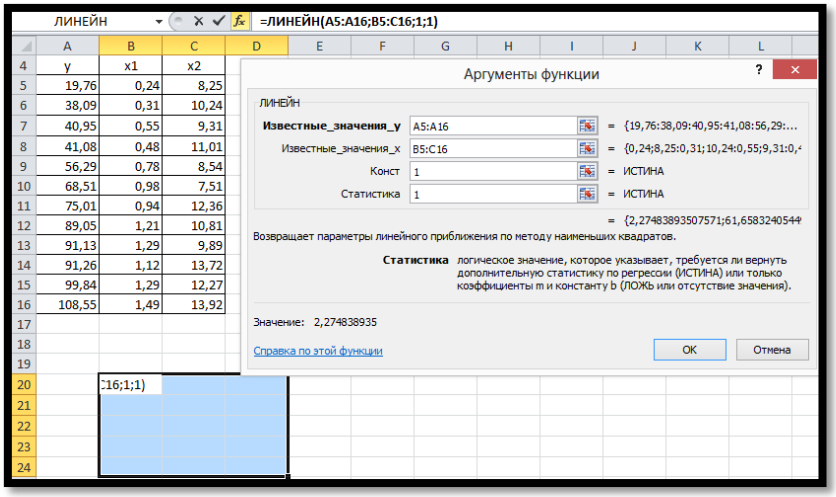
	A	B	C
1			
2			
3			
4	y	x1	x2
5	19,76	0,24	8,25
6	38,09	0,31	10,24
7	40,95	0,55	9,31
8	41,08	0,48	11,01
9	56,29	0,78	8,54
10	68,51	0,98	7,51
11	75,01	0,94	12,36
12	89,05	1,21	10,81
13	91,13	1,29	9,89
14	91,26	1,12	13,72
15	99,84	1,29	12,27
16	108,55	1,49	13,92

Для автоматизации расчетов воспользуемся встроенной функцией «Линейн».

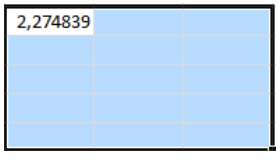
Выделяем область 3×5, открываем меню «формулы» – «другие функции» – «статистические» – «Линейн».



В появившемся окне заполняем поля

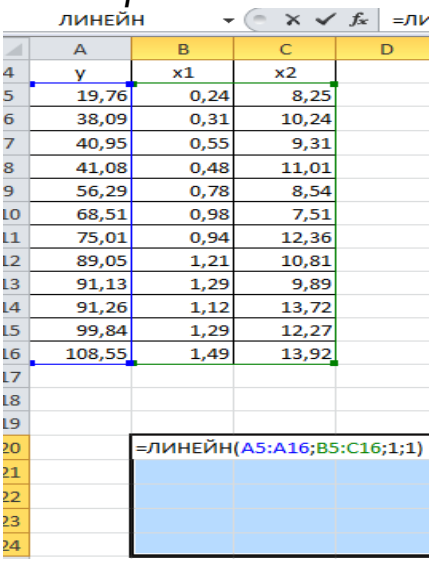


Нажимаем ОК



В выделенной области будет одно число последнего параметра регрессии.

Далее нажимаем *f4* и *ctrl + shift + Enter*.



Получим набор всех решений, необходимых для решения и анализа регрессионного уравнения.

2,274839	61,65832	-10,8153
0,583279	2,944476	5,364618
0,988422	3,410306	#Н/Д
384,1781	9	#Н/Д
8936,13	104,6717	#Н/Д

Для удобства восприятия полученной таблицы решений распишем в какой ячейке какие решения получены.

$\widehat{b}_2$	$\widehat{b}_1$	a
m_{b2}	m_{b1}	m_a
R^2	S	*#Н/Д*
F	df	*#Н/Д*
Q_r	Q_e	*#Н/Д*

Используя данные первой строки, построим модель вида:

$$\widehat{y} = -10.82 + 61.66x_1 + 2.27x_2$$

Запишем данные в столбце таблицы *y(t)*.

Добавим для начала столбец в таблицу и в первой ячейке запишем формулу расчета теоретических значений годового товарооборота:

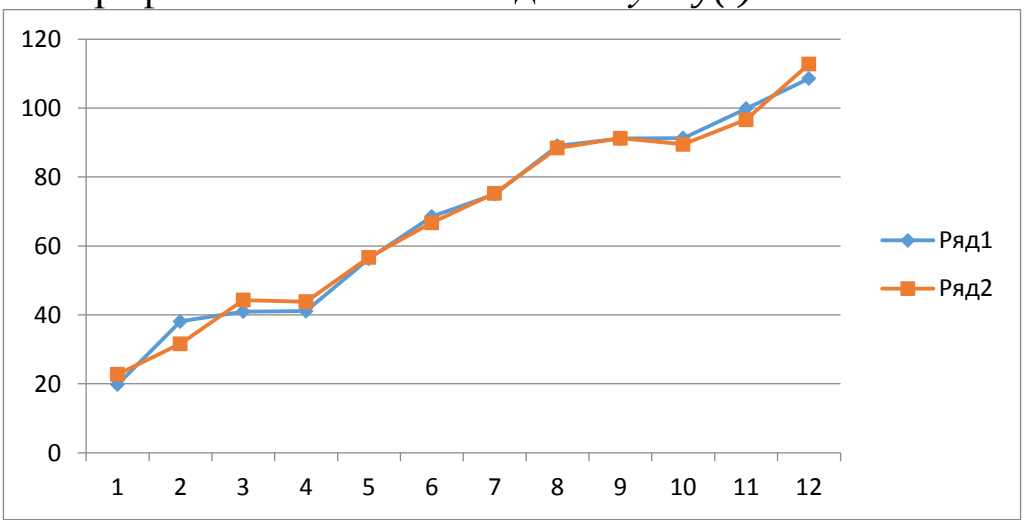
= $D$20 + $C$20*B5 + $B$20*C5

Заметим, что значения коэффициентов регрессии закреплены на абсолютном уровне.

Совершаем протягивание ячеек для расчета оставшихся значений столбца *y(t)*

y	x1	x2	y(t)
19,76	0,24	8,25	22,75
38,09	0,31	10,24	31,59
40,95	0,55	9,31	44,28
41,08	0,48	11,01	43,83
56,29	0,78	8,54	56,71
68,51	0,98	7,51	66,69
75,01	0,94	12,36	75,26
89,05	1,21	10,81	88,38
91,13	1,29	9,89	91,22
91,26	1,12	13,72	89,45
99,84	1,29	12,27	96,64
108,55	1,49	13,92	112,72

Изобразим графически линию поведения *y* и *y(t)*



Линия регрессии достаточно хорошо описывает поведение годового товарооборота.

Коэффициент корреляции.

Индекс множественного коэффициента корреляции можно определить при помощи решений, полученных при использовании функции «Линейн».

$\widehat{b}_2$	$\widehat{b}_1$	a
m_{b2}	m_{b1}	m_a
R^2	S	#Н/Д
F	df	#Н/Д
Q_r	Q_e	#Н/Д

Извлекая корень квадратный из значения коэффициента детерминации, найдем коэффициент корреляции.

2,27	61,66	-10,82
0,583279	2,944476	5,364618
0,988422	3,410306	#Н/Д
384,1781	9	#Н/Д
8936,13	104,6717	#Н/Д

$R = \sqrt{R^2} = \sqrt{0{,}9884} = 0{,}994$

Коэффициент корреляции равен 0,99, что говорит об очень тесной зависимости годового товарооборота (*y*) от совокупного влияния факторов *x*.

Коэффициент детерминации

Коэффициент детерминации можно определить как квадрат множественного индекса корреляции.

$$\boldsymbol{R = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2}}$$

R = 0, 98

Коэффициент детерминации показывает, что результат *y* на 98% зависит от совокупного влияния факторов *X*. Т.е., годовой товарооборот на *98%* зависит от торговой площади ($\boldsymbol{X_1}$) и среднего числа посетителей в день($\boldsymbol{X_2}$).

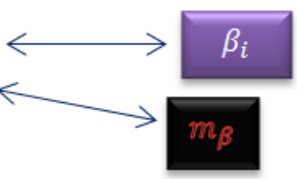
Оценка статистической значимости коэффициентов регрессии

Статистическую оценку значимости коэффициентов множественной регрессии можно определить, разделив значение соответствующего параметра уравнения регрессии на его стандартную ошибку:

$$t_\beta = \frac{\beta_i}{m_\beta}$$

Эти значения получены в первой и второй строках выделенной нами области

2,27	61,66	-10,82
0,583279	2,944476	5,364618
0,988422	3,410306	#Н/Д
384,1781	9	#Н/Д
8936,13	104,6717	#Н/Д



При расчете статистики параметра a будем брать его по модулю

$$t_a = \frac{a}{m_a} = \frac{10.82}{5,36} = 2,02$$

$$t_{b1} = \frac{b_1}{m_{b1}} = \frac{61,66}{2,94} = 20,94$$

$$t_{b2} = \frac{b_2}{m_{b2}} = \frac{2,27}{0,58} = 3,9$$

$t_T = 2,62$

$t_a < t_T \rightarrow H_0$

$t_{b1} > t_T \rightarrow H_1$

$t_{b1} > t_T \rightarrow H_1$

Оценка значимости уравнения регрессии.

Для оценки значимости уравнения регрессии воспользуемся решениями, полученными ранее:

2,27	61,66	-10,82
0,583279	2,944476	5,364618
0,988422	3,410306	#Н/Д
384,1781	9	#Н/Д
8936,13	104,6717	#Н/Д

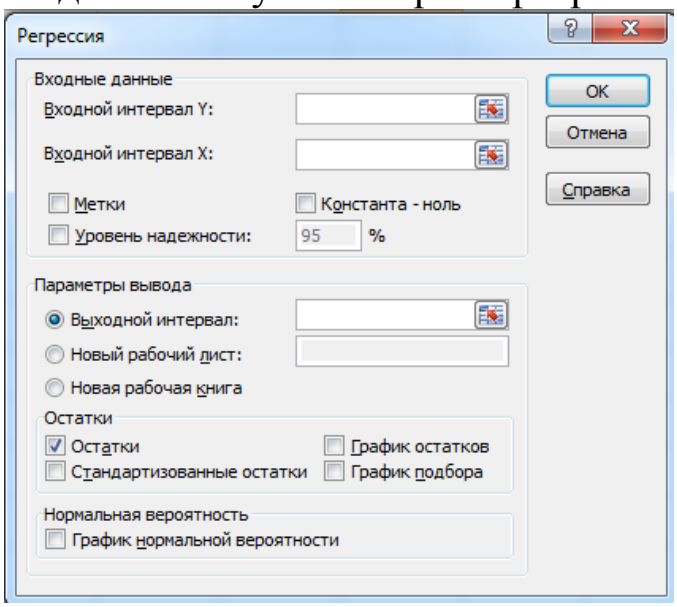
F– статистика Фишера для оценки значимости уравнения регрессии равна 384,18. Сравнивая это значение с табличным значением *F*-критерия Фишера (F_T = 4,26) на уровне значимости a = 0,05 при степени свободы k_2 = n– m-1 = 12 – 2 -1 = 9 и количестве регрессоров k_1 = 2, ($384,18 > 4,26 \rightarrow H_1$), принимается гипотеза о значимости уравнения регрессии.

3.1. Расчеты с помощью инструментов «Регрессия», «Пакет анализа»

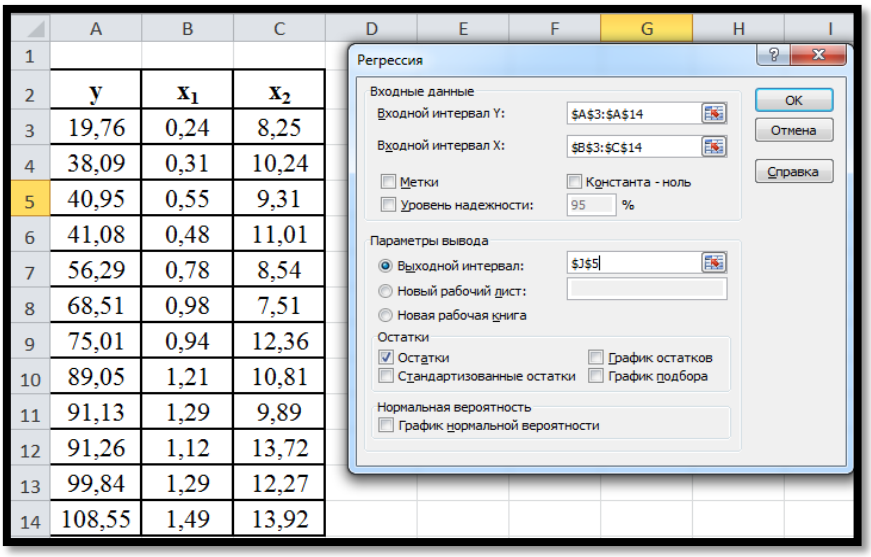
Инструмент «Регрессия», как и функция ЛИНЕЙН, может работать не только с одной объясняющей переменной, но и с несколькими.

При применении инструмента «Пакет анализа» в случае множественной регрессии особенностью является только то, что при вводе входного массива данных объясняющих переменных

необходимо задать диапазон ячеек нескольких соседних столбцов. Заполняем поля ввода как в случае с парной регрессией:



Если необходимые столбцы не соседние, их надо такими сделать путём замены столбцов местами. В данной задаче необходимо задать входной интервал X «B3:C14» в окне, аналогичном изображённому на рисунке.



	y	x_1	x_2
3	19,76	0,24	8,25
4	38,09	0,31	10,24
5	40,95	0,55	9,31
6	41,08	0,48	11,01
7	56,29	0,78	8,54
8	68,51	0,98	7,51
9	75,01	0,94	12,36
10	89,05	1,21	10,81
11	91,13	1,29	9,89
12	91,26	1,12	13,72
13	99,84	1,29	12,27
14	108,55	1,49	13,92

Поставим галочку в поле «Остатки» формы исходных данных инструмента «Регрессия». Закажем вывод результатов вычислений на отдельном листе. Для этого в окне переключатель "Параметры вывода" должен стоять в положении "Новый рабочий лист". Если же вывод результатов вычислений мы хотим увидеть на этом же листе , тогда в поле «Выходной интервал» нужно указать ячейку, с которой и будет начинаться вывод результатов вычислений. Сравниваем полученные данные с вычисленными выше функцией ЛИНЕЙН.

Регрессионная статистика	
Множественный R	0,994
R-квадрат	0,988
Нормированный R-квадрат	0,986
Стандартная ошибка	3,410
Наблюдения	12

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	2	8936,1298	4468,0649	384,1781	0,00000000
Остаток	9	104,6717	11,6302		
Итого	11	9040,8015			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	-10,8153	5,3646	-2,0160	0,0746	-22,9509	1,3203
Переменная X 1	61,6583	2,9445	20,9403	0,0000	54,9975	68,3192
Переменная X 2	2,2748	0,5833	3,9001	0,0036	0,9554	3,5943

Здесь приведены все рассмотренные выше расчеты с соответствующими пояснениями.

Галочка на вкладке вывод остатков дает нам результаты предсказанных ($\hat{y}$)наблюдений, а также отклонения между фактическими наблюдениями и теоретически полученными по регрессионной модели.

Наблюдение	*Предсказанное Y*	*Остатки*
1	22,7501	-2,9901
2	31,5931	6,4969
3	44,2755	-3,3255
4	43,8267	-2,7467
5	56,7053	-0,4153
6	66,6939	1,8161
7	75,2605	-0,2505
8	88,3823	0,6677
9	91,2221	-0,0921
10	89,4528	1,8072
11	96,6362	3,2038
12	112,7214	-4,1714

Таким образом, основная нагрузка, связанная с трудоемким процессом вычислений, которая ложилась на студента, может быть легко преодолена одним лишь желанием научиться чему – нибудь новому.

ЗАКЛЮЧЕНИЕ

Приведенные в пособии примеры являются простейшими и рассчитаны на неискушенного слушателя, впервые столкнувшегося с методологическим аппаратом эконометрики. Малое число наблюдений в заданиях позволяет практически «вручную» (на калькуляторе) выполнить необходимые расчеты. Поставленные задания носят скорее технический характер, хотя приведенные в пособии примеры связаны с реальной экономикой. На начальном этапе изучения эконометрики эта работа необходима. Слушатель учится правильно читать листинги эконометрических отчетов, понимает, как получена та или иная характеристика модели. Эта работа делает более доступным лекционный материал.

Использование эконометрического пакета позволит существенно поднять планку заданий и максимально приблизить их к практике современных эконометрических исследований.

Библиографический список

1. Айвазян С.А. Прикладная статистика. Основы эконометрики: учебник для вузов: В 2 т. / Айвазян С.А., Мхитарян В.С.; С.А. Айвазян, В.С. Мхитарян. – 2-е изд., испр. – М.: ЮНИТИ, 2001. – Т. 1: Теория вероятностей и прикладная статистика. – 656 с.
2. Берндт, Э.Р. Практика эконометрики: классика и современность: учебник / Э.Р. Берндт. – М.: ЮНИТИ-ДАНА, 2005. – 863 с.
3. Доугерти, К. Введение в эконометрику / К. Доугерти. – М.: ИНФРА-М, 1997. – 402 с.
4. Елисеева, И.И. Эконометрика: учебное пособие / И.И. Елисеева, С. В. Курышева, Д.М. Гордиенко и др. – М.: Финансы и статистика, 2001.
5. Замков О.О. Эконометрические методы в макроэкономическом анализе: курс лекций / О.О. Замков. – М.: ГУ ВШЭ, 2001. – 122 с.
6. Кремер Н.Ш. Эконометрика / Н.Ш. Кремер, Б.А. Путко. – М.: ЮНИТИ, 2005. – 311 с.
7. Магнус Я.Р. Эконометрика. Начальный курс / Я.Р. Магнус, П. К. Катышев, А.А. Пересецкий. – М.: Дело, 1997. – 247с.
8. Практикум по эконометрике: учебное пособие / под ред. И.И. Елисеевой. – М.: Финансы и статистика, 2002. – 191с.

Интернет-ресурсы

9. Колеников С. О. Прикладной эконометрический анализ в статистическом пакете Stata6 в формате PDF. – Режим доступа: http://www.komkon.org/~ tacik/Stata6Ec.pdf.
10. Математическое Бюро. Учебники по эконометрике и статистике. – Режим доступа: http://www.matburo.ru/st_subject.php?p=ec.
11. Образовательный математический сайт. – Режим доступа: http://www.exponenta.ru.
12. Эконометрика. Библиотека. Единое окно доступа к образовательным ресурсам. – Режим доступа: http://window.edu.ru/window/library?p_rubr =2.2.76.4.8

СТАТИСТИКО-МАТЕМАТИЧЕСКИЕ ТАБЛИЦЫ

Приложение 1.
Таблица значений F-критерия Фишера на уровне значимости α = 0,05

k_2 \ k_1	1	2	3	4	5	6	8	12	24	∞
1	161,45	199,50	215,72	224,57	230,17	233,97	238,89	243,91	249,04	254,32
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,45	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,64	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,77	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	3,84	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,12	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,07	2,90	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,91	2,74	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,79	2,61	2,40
12	4,75	3,88	3,49	3,26	3,11	3,00	2,85	2,69	2,50	2,30
13	4,67	3,80	3,41	3,18	3,02	2,92	2,77	2,60	2,42	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,53	2,35	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,48	2,29	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,42	2,24	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,38	2,19	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,34	2,15	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,31	2,11	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,28	2,08	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,42	2,25	2,05	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2,23	2,03	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,38	2,20	2,00	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,36	2,18	1,98	1,73
25	4,24	3,38	2,99	2,76	2,60	2,49	2,34	2,16	1,96	1,71
26	4,22	3,37	2,98	2,74	2,59	2,47	2,32	2,15	1,95	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,30	2,13	1,93	1,67
28	4,20	3,34	2,95	2,71	2,56	2,44	2,29	2,12	1,91	1,65
29	4,18	3,33	2,93	2,70	2,54	2,43	2,28	2,10	1,90	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,09	1,89	1,62
35	4,12	3,26	2,87	2,64	2,48	2,37	2,22	2,04	1,83	1,57
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,00	1,79	1,51
45	4,06	3,21	2,81	2,58	2,42	2,31	2,15	1,97	1,76	1,48

Приложение 2.

Критические значения t-критерия Стьюдента при уровне значимости 0,05 и 0,01.

Число степеней свободы	α=0,05	α=0,01	Число степеней свободы	α=0,05	α=0,01
1	12,706	63,657	16	2,1199	2,9208
2	4,3027	9,9248	17	2,1098	2,8982
3	3,1825	5,8409	18	2,1009	2,8784
4	2,7764	4,6041	19	2,093	2,8609
5	2,5706	4,0321	20	2,086	2,8453
6	2,4469	3,7074	21	2,0796	2,8314
7	2,3646	3,4995	22	2,0739	2,8188
8	2,3060	3,3554	23	2,0687	2,8073
9	2,2622	3,2498	24	2,0639	2,7969
10	2,2281	3,1693	25	2,0595	2,7874
11	2,2010	3,1058	26	2,0555	2,7787
12	2,1788	3,0545	27	2,0518	2,7707
13	2,1604	3,0123	28	2,0484	2,7633
14	2,1448	2,9768	29	2,0452	2,7564
15	2,1315	2,9467	30	2,0423	2,75

Приложение 3.

Значения статистик Дарбина – Уотсона при 5% -ном уровне значимости.

n	k=1		k=2	
	d_l	d_u	d_l	d_u
6	0.61	1.4	-	-
7	0.7	1.36	0.47	1.90
8	0.76	1.33	0.56	1.78
9	0.82	1.32	0.63	1.7
10	0.88	1.32	0.7	1.64
11	0.93	1.32	0.66	1.6
12	0.97	1.33	0.81	1.58
13	1.01	1.34	0.86	1.56
14	1.05	1.35	0.91	1.55
16	1.1	1.37	0.98	1.54
17	1.13	1.38	1.02	1.54
18	1.16	1.39	1.05	1.53
19	1.18	1.4	1.08	1.53
20	1.2	1.41	1.1	1.54

Подписано в печать 31.12.2015 г. Формат 60х90 1/6
У.п.л. 2.19. Бумага офисная. Печать-ризография.
Тираж 100 экз.

Издательство Чеченского государственного университета
Адрес: 364037 ЧР, г. Грозный,
ул. Киевская, 33.