Курс лекций соответствует учебной программе по дисциплине «Эконометрика»

Раздел №1. Определение эконометрики

1. Предмет эконометрики.

В одном из наиболее распространенных в России вводном курсе западной экономической теории сказано: "Статистический анализ экономических данных называется эконометрикой, что буквально означает: наука об экономических измерениях". Действительно, термин "эконометрика" состоит из двух частей: "эконо-" - от "экономика" и "-метрика" - от "измерение". Эконометрика (в другом русско- и англоязычном варианте названия этой дисциплины - эконометрия) входит в обширное семейство дисциплин, посвященных измерениям и применению статистических методов в различных областях науки и практики. К этому семейству относятся, в частности, биометрика (или биометрия), технометрика, наукометрия, психометрика, хемометрика (наука об измерениях и применении статистических методов в химии). Особняком стоит социометрия - этот термин закрепился за статистическими методами анализа взаимоотношений в малых группах, т.е. за небольшой частью такой дисциплины, как статистический анализ в социологии. Эконометрика, как и другие "метрики", посвящена развитию и применению статистических методов в конкретной области науки и практики - в экономике, прежде всего в теории и практике менеджмента.

В мировой науке эконометрика занимает достойное место. Нобелевские премии по экономике получили эконометрики Ян Тильберген, Рагнар Фриш, Лоуренс Клейн, Трюгве Хаавельмо. В 2000 г. к ним добавились еще двое - Джеймс Хекман и Дэниель Мак-Фадден. Выпускается ряд научных журналов, полностью посвященных эконометрике, в том числе:

Journal of Econometrics (Швеция),
Econometric Reviews (США),
Econometrica (США),
Sankhya. Indian Journal of Statistics. Ser.D. Quantitative Economics (Индия),
Publications Econometriques (Франция).

Однако в нашей стране по ряду причин эконометрика не была сформирована как самостоятельное направление научной и практической деятельности, в отличие, например, от Польши, которая стараниями О.Ланге и его коллег покрыта сетью эконометрических "институтов" (в российской терминологии - кафедр вузов). В настоящее время в России начинают развертываться эконометрические исследования, в частности, начинается широкое преподавание этой дисциплины.

2. Особенности эконометрического метода.

Кратко рассмотрим современную структуру эконометрики. Знакомство с ней необходимо для обоснованных суждений о возможностях применения эконометрических методов и моделей в экономических и технико-экономических исследованиях

В эконометрике, как дисциплине на стыке экономики (включая менеджмент) и статистического анализа, естественно выделить три вида научной и прикладной деятельности (по степени специфичности методов, сопряженной с погруженностью в конкретные проблемы):

а) разработка и исследование эконометрических методов (методов прикладной статистики) с учетом специфики экономических данных;

б) разработка и исследование эконометрических моделей в соответствии с конкретными потребностями экономической науки и практики;

в) применение эконометрических методов и моделей для статистического анализа конкретных экономических данных.

Кратко рассмотрим три только что выделенных вида научной и прикладной деятельности. По мере движения от а) к в) сужается широта области применения конкретного эконометрического метода, но при этом повышается его значение для анализа конкретной экономической ситуации. Если работам вида а) соответствуют научные результаты, значимость которых оценивается по общеэконометрическим критериям, то для работ вида в) основное - успешное решение задач конкретной области экономики. Работы вида б) занимают промежуточное положение, поскольку, с одной стороны, теоретическое изучение эконометрических моделей может быть весьма сложным и математизированным (см., например, монографию [5]), с другой - результаты представляют интерес не для всей экономической науки, а лишь для некоторого направления в ней.

Прикладная статистика - другая область знаний, чем математическая статистика. Курс математической статистики состоит в основном из доказательств теорем. В курсах прикладной статистики и эконометрики основное - методология анализа данных и алгоритмы расчетов, а теоремы приводятся как обоснования этих алгоритмов, доказательства же, как правило, опускаются (их можно найти в научной литературе).

Математическая статистика играет роль математического фундамента для прикладной статистики. Прикладная статистика нацелена на решение реальных задач. Поэтому в ней возникают новые постановки математических задач анализа статистических данных, развиваются и обосновываются новые методы. Обоснование часто проводится математическими методами, т.е. путем доказательства теорем. Большую роль играет методологическая составляющая - как именно ставить задачи, какие предположения принять с целью дальнейшего математического изучения. Велика роль современных информационных технологий, в частности, компьютерного эксперимента.

Подведем итог. Хотя статистические данные собираются и анализируются с незапамятных времен (см., например, Книгу Чисел в Ветхом Завете), современная математическая статистика как наука была создана, по общему мнению специалистов, сравнительно недавно - в первой половине XX в. Именно тогда были разработаны основные идеи и получены результаты, излагаемые ныне в учебных курсах математической статистики. После чего специалисты по математической статистике занялись внутриматематическими проблемами, а для теоретического обслуживания проблем практического анализа статистических данных стала формироваться новая дисциплина - прикладная статистика

3. Измерения в эконометрике.

В настоящее время статистическая обработка данных проводится, как правило, с помощью соответствующих программных продуктов. Разрыв между математической и прикладной статистикой проявляется, в частности, в том, что большинство методов, включенных в статистические пакеты программ (например, в заслуженные Statgraphics и SPSS или в более новую систему Statistica), даже не упоминается в учебниках по математической статистике. В результате специалист по математической статистике оказывается зачастую беспомощным при обработке реальных данных, а пакеты программ применяют (что еще хуже - и разрабатывают) лица, не имеющие необходимой теоретической подготовки. Естественно, что они допускают разнообразные ошибки, в том числе в таких ответственных документах, как государственные стандарты по статистическим методам (ниже подробнее рассказано об удручающих результатах анализа этих стандартов; итоги суммированы в статье).

Для анализа экономических данных могут применяться все разделы прикладной статистики, а именно:

- статистика случайных величин;
- многомерный статистический анализ;
- статистика временных рядов и случайных процессов;
- статистика объектов нечисловой природы, в том числе статистика интервальных данных.

Перечисленные четыре области выделены на основе математической природы элементов выборки: в первой из них это - числа, во второй - вектора, в третьей - функции, в четвертой - объекты нечисловой природы, т.е. элементы пространств, в которых нет операций сложения и умножения на число. Примерами объектов нечисловой природы являются значения качественных признаков, бинарные отношения (ранжировки, разбиения, толерантности), последовательности из 0 и 1, множества, нечеткие множества, интервалы, тексты (см. главы 8 и 9 ниже)..

Как и для применений статистических методов в иных областях, в эконометрике решаются задачи описания данных (в том числе усреднения), оценивания, проверки гипотез, восстановления зависимостей, классификации объектов и признаков, прогнозирования, принятия статистических решений и др.

Однако в некоторых отношениях экономические данные отличаются от технических или астрономических, и эти отличия необходимо учитывать при выборе методов анализа конкретных экономических данных.

Многие экономические показатели неотрицательны. Значит, их надо описывать неотрицательными случайными величинами. А вот нормальные распределения принципиально не подходят, поскольку для них вероятность отрицательных значений всегда положительна.

Экономические процессы развиваются во времени, поэтому большое место в эконометрике занимают вопросы анализа и прогнозирования временных рядов, в том числе многомерных. При этом в одних задачах больше внимания уделяют изучению трендов (средних значений, математических ожиданий), например, при анализе динамики цен. В других же - важны отклонения от средней тенденции, например, при применении контрольных карт (карт Шухарта, кумулятивных сумм и др.). Однако в целом спектральный анализ и выделение различных периодов, циклов и типов волн менее распространены, чем, скажем, в биометрике и медицине.

В экономике доля нечисловых данных существенно выше, чем в технике и технологии, соответственно больше применений для статистики объектов нечисловой природы (ниже разберем это утверждение подробнее).

Количество изучаемых объектов в экономическом исследовании часто ограничено в принципе, поэтому обоснование вероятностных моделей в ряде случаев затруднено. Уникальные объекты, например, город Москва, трудно рассматривать как элемент выборки из генеральной совокупности с каким-то определенным распределением, поскольку подобное рассмотрение противоречит здравому смыслу. Вспоминается давняя обложка журнала "Крокодил", на которой изображены два хозяйственника с монетой в руках: "Если упадет орлом, будем строить завод, если решкой - не будем". Подобная рандомизация решений выглядит бессмысленной при принятии ровно одного решения, однако при контроле качества в массовом производстве такой подход оправдан.

Поэтому в эконометрике часто применяются детерминированные методы анализа данных, в отличие от, например, технических наук, в которых обычным является использование вероятностных моделей. Неопределенность приходится описывать не в терминах вероятностно-статистических моделей, а иными способами, например, в терминах теории нечеткости (fuzzy sets theory) или математики и статистики интервальных данных.

Итак, специфика эконометрики проявляется не в перечне применяемых для анализа конкретных экономических данных статистических методов, а в частоте использования тех или иных методов.

Раздел №2. Корреляционная зависимость.

Регрессия и корреляция широко используется при анализе связей между явлениями. Прежде всего, в экономике – исследование зависимости объемов производства от целого ряда факторов: размера основных фондов, обеспеченности предприятия квалифицированным персоналом и других; зависимости спроса или потребления населения от уровня дохода, цен на товары и т.д. Экономические показатели являются многомерными случайными величинами.

В большинстве случаев между переменными, характеризующими экономические величины, существуют зависимости, отличающиеся от функциональных. Она возникает, когда один из факторов зависит не только от другого, но и от ряда случайных условий, оказывающих влияние на один или оба фактора. В этом случае ее называют стохастической **(корреляционной)** и говорят, что переменные **коррелируют**. Виды стохастических связей между факторами могут быть линейными и нелинейными, положительными или отрицательными. Возможна такая ситуация, когда между факторами невозможно установить какую–либо зависимость.

Однако при изучении влияния одного явления на другое удобно работать именно с функциями, связывающими эти явления. Задачи построения функциональной зависимости между факторами, анализа полученных результатов и прогнозирования решаются с помощью **регрессионного анализа**.

Экономические данные представляют собой количественные характеристики каких-либо экономических объектов или процессов. Они формируются под действием множества факторов, не все из которых доступны внешнему контролю. Неконтролируемые факторы могут принимать случайные значения из некоторого множества значений и тем самым обуславливать случайность данных, которые они определяют. Стохастическая (вероятностная) природа экономических данных обуславливает необходимость применения соответствующих статистических методов для их обработки и анализа.

Статистические распределения характеризуются наличием более или менее значительной вариации в величине признака у отдельных единиц совокупности. Естественно, возникает вопрос о том, какие же причины формируют уровень признака в данной совокупности и каков конкретный вклад каждой из них. Изучение зависимости вариации признака от окружающих условий и составляет содержание теории корреляции.

Изучение действительности показывает, что вариация каждого изучаемого признака находится в тесной связи и взаимодействии с вариацией других признаков, характеризующих исследуемую совокупность единиц. Вариация уровня производительности труда работников предприятий зависит от степени совершенства применяемого оборудования, технологии, организации производства, труда и управления и других самых различных факторов.

Основная задача **корреляционного анализа** заключается

- ☑ в выявлении взаимосвязи между случайными переменными путем точечной и интервальной оценки парных (частных) коэффициентов корреляции,
- ☑ вычисления и проверки значимости множественных коэффициентов корреляции и детерминации
- ☑ отбор факторов, оказывающих наиболее существенное влияние на результативный признак, на основании измерения степени связи между ними; обнаружение ранее неизвестных причинных связей.

Корреляция непосредственно не выявляет причинных связей между параметрами, но устанавливает численное значение этих связей и достоверность суждений об их наличии.

При проведении корреляционного анализа вся совокупность данных рассматривается как множество переменных (факторов), каждая из которых содержит n – наблюдений; x_{ik} – i-ое наблюдение k-ой переменной. Основными средствами анализа данных являются парные коэффициенты корреляции, частные коэффициенты корреляции и множественные коэффициенты корреляции.

Рассматривая зависимости между признаками, необходимо выделить, прежде всего, две категории зависимости: 1) функциональные и 2) корреляционные.

Функциональные связи характеризуются полным соответствием между изменением факторного признака и изменением результативной величины, и каждому значению признака-фактора соответствуют вполне определенные значения результативного признака. Функциональная зависимость может связывать результативный признак с одним или несколькими факторными признаками. Так, величина начисленной заработной платы при повременной оплате труда зависит от количества отработанных часов.

В корреляционных связях между изменением факторного и результативного признака нет полного соответствия, воздействие отдельных факторов проявляется лишь в среднем при массовом наблюдении фактических данных. Одновременное воздействие на изучаемый признак большого количества самых разнообразных факторов приводит к тому, что одному и тому же значению признака-фактора соответствует целое распределение значений результативного признака, поскольку в каждом конкретном случае прочие факторные признаки могут изменять силу и направленность своего воздействия.

При сравнении функциональных и корреляционных зависимостей следует иметь в виду, что при наличии функциональной зависимости между признаками можно, зная величину факторного признака, точно определить величину результативного признака. При наличии же корреляционной зависимости устанавливается лишь тенденция изменения результативного признака при изменении величины факторного признака. В отличие от жесткости функциональной связи корреляционные связи характеризуются множеством причин и следствий и устанавливаются лишь их тенденции.

Числовые характеристики системы двух случайных величин

Две случайные величины называются независимыми, если закон распределения одной из них не зависит от того, какие возможные значения приняла другая величина. Из этого определения следует, что условные распределения независимых величин равны их безусловным распределениям. Укажем необходимые и достаточные условия независимости случайных величин.

Теорема 4.1. Для того чтобы случайные величины X и Y были независимыми, необходимо и достаточно, чтобы функция распределения системы (X, Y) была равна произведению функций распределения составляющих:

$$F(x, y) = F_1(x)F_2(y).$$

Теорема 4.2. Для того чтобы непрерывные случайные величины X и Y были независимыми, необходимо и достаточно, чтобы плотность вероятности системы (X, Y) была равна произведению плотностей вероятностей составляющих:

$$f(x, y) = f_1(x)f_2(y).$$

Для описания системы двух случайных величин кроме математических ожиданий и дисперсий составляющих используют и другие характеристики, к которым относятся корреляционный момент и коэффициент корреляции.

Корреляционным моментом μ_{xy} случайных величин X и Y называют математическое ожидание произведения отклонений этих величин:

$$\mu_{xy} = M((X - M(X))(Y - M(Y))).$$

Для вычисления корреляционного момента дискретных величин используют формулу

$$\mu_{xy} = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_i - M(X))(y_j - M(Y))P(x_i, y_j).$$

а для непрерывных величин

Корреляционный момент служит для характеристики связи между величинами X и Y.

Теорема 4.3. Корреляционный момент двух независимых случайных величин X и Y равен нулю.

Из теоремы 4.3 следует, что если корреляционный момент двух случайных величин X и Y не равен нулю, то X и Y — зависимые случайные величины.

Коэффициентом корреляции r_{xy} случайных величин X и Y называют отношение корреляционного момента к произведению средних квадратических отклонений этих величин:

$$r_{xy} = \frac{\mu_{xu}}{\sigma_x \sigma_y}.$$

Очевидно, коэффициент корреляции двух независимых случайных величин равен нулю (так как $\mu_{xy} = 0$).

Две случайные величины X и Y называют коррелированными, если их корреляционный момент (или коэффициент корреляции) отличен от нуля; X и Y называют ***некоррелированными*** величинами, если их корреляционный момент равен нулю.

Две коррелированные величины также и зависимы. Обратное утверждение не всегда имеет место, т. е. если две величины зависимы, то они могут быть как коррелированными, так и некоррелированными. Другими словами, корреляционный момент двух зависимых величин может быть не равен нулю, но может и равняться нулю.

Парная линейная регрессия.

Дает возможность провести более полный анализ, полученного уравнения линии тренда с использованием методов математической статистики.

Коэффициенты уравнения линейной регрессии находятся по выборочным данным и являются величинами случайными, поэтому надо провести анализ их значимости (существенности). Надо определить значимость всего уравнения регрессии и самое главное прогноз по построенному уравнению и оценку его значимости.

При построении линейного тренда принималось предположение о том, что **линейная модель** наилучшим образом характеризует зависимость между x и y:

$$y=\beta_0+\beta_1 x+\varepsilon_i, \qquad 4.1$$

где β_0 и β_1 **параметры** модели; ε – случайная величина (**возмущение**), характеризующая влияние неучтенных факторов.

Уравнение прямой (4.2), коэффициенты которого находят по выборочным данным, называют уравнением **регрессии** и обозначают **ŷ**:

$$\hat{y}=b_0 + b_1 \cdot x, \qquad 4.2$$

Коэффициенты b_0 и b_1 регрессии находятся по методу наименьших квадратов и являются только **оценками** параметров модели (соответственно β_0 и β_1). Для получения **наилучших оценок** необходимо, чтобы выполнялся ряд предпосылок **относительно случайного отклонения** $e_i=y_i-\hat{y}_i=y_i-b_0-b_1x_i$ индекс i означает значение факторов в одноименном испытании. Это условия Гаусса–Маркова (Приложение 1), а так же предположения:

- случайные отклонения имеют нормальный закон распределения;
- отсутствуют ошибки спецификации;
- число наблюдений достаточно большое: как минимум в шесть раз превышает число объясняющих факторов и другие.

Оценку b_1 называют **коэффициентом регрессии**. Ее значение показывает среднее изменение результата y с изменением фактора x на одну единицу.

Тесноту связи между x и y оценивает коэффициент линейной корреляции r_{xy} Можно установить зависимость между коэффициентом регрессии и коэффициентом корреляции:

$$r_{xy} = b_1 \frac{\sigma_x}{\sigma_y} = \frac{\overline{yx} - \bar{y} \cdot \bar{x}}{\sigma_x \sigma_y} = \frac{\operatorname{cov}(x, y)}{\sigma_x \sigma_y}. \quad 4.3$$

В качестве меры рассеивания **фактического** значения у относительно **теоретического** значения $\hat{y}$ (находится по уравнению регрессии) используется **стандартная ошибка** уравнения регрессии, которая определяется по формуле:

$$S = \sqrt{\frac{\sum e_i^2}{n-2}}. \quad 4.4$$

ОЦЕНКА КАЧЕСТВА ПОЛУЧЕННОГО УРАВНЕНИЯ РЕГРЕССИИ СОДЕРЖИТ СЛЕДУЮЩИЕ ПУНКТЫ:

- Оценка значимости коэффициентов регрессии;
- Построение доверительных интервалов для каждого коэффициента;
- Оценка значимости всего уравнения регрессии;
- Построение прогнозного значения и доверительного интервала к ним.

Для определения **статистической значимости** коэффициентов регрессии и корреляции необходимо рассчитать **t-статистики** Стьюдента лучше всего это сделать с помощью встроенной функции СТЬДРАСПОБР.

ОЦЕНКА ЗНАЧИМОСТИ КОЭФФИЦИЕНТОВ РЕГРЕССИИ И КОРРЕЛЯЦИИ

Устанавливает надежность полученных результатов. Случайные ошибки коэффициента корреляции и оценок параметров линейной модели вычисляются по формулам:

$$S_1 = S_{b_1} = \sqrt{\frac{S^2}{\sum_i (x_i - \bar{x})^2}} = \frac{S}{\sqrt{n} \cdot \sigma_x}$$ **– стандартное отклонение коэффициента b_1.** 4.5

$$S_0 = S_{b_0} = \sqrt{S_1^2 \cdot \frac{\sum_i x_i^2}{n}} = \sqrt{S_1^2 \cdot \overline{x^2}} = S_1 \cdot \sqrt{\overline{x^2}}$$ – **стандартное отклонение коэффициента b_0.** **4**.6

$$S_r = \sqrt{\frac{1 - r_{xy}^2}{n-2}}.$$ – **стандартное отклонение коэффициента корреляции.** 4.7

Любое стандартное отклонение иногда называют **стандартной ошибкой** соответствующего коэффициента.

Рассматривается основная **гипотеза о равенстве параметров регрессии нулю**.

H_0: $b_i = 0$ – коэффициент незначим;

H_1: $b_i \neq 0$ – коэффициент значимый

По выборке находят t-статистики ($T_{набл}$):

$$t_1 = T_{набл}(b_1) = \frac{b_1}{S_{b_1}}, \quad t_0 = T_{набл}(b_0) = \frac{b_0}{S_{b_0}}, \quad t_r = T_{набл}(r_{xy}) = \frac{r_{xy}}{S_r}. \quad 4.8$$

Критическое значение $T_{кр}$ для t-статистик находят с помощью распределения Стьюдента. Для этого надо знать **объем** выборки и задать **уровень значимости** α. Например, для α=0,05 и n=14, $\mathbf{T_{кр} = t_{\alpha/2,\, n-2}} = t_{0,025,\, 12} = 2{,}179$.

Выдвинутая гипотеза:

- **принимается, если выполняется неравенство** $|T_{набл}| < T_{кр}$ и делают вывод, что коэффициент незначим (равен нулю);
- **отвергается, если** $|T_{набл}| > T_{кр}$ и делают вывод, что коэффициент значим.

Более подробно о проверке гипотез можно прочитать в разделе !!! часть 1.

Часто при проверке качества коэффициентов используют «**грубое прав**ило»:

- если $|t| \leq 1$ ($b_j < S_j$), то коэффициент статистически незначим;

- если $1<|t|\le 2$ ($b_j<2S_j$), то коэффициент относительно слабо значим, рекомендуется воспользоваться таблицей критических точек распределения Стьюдента;
- если $2<|t|\le 3$, то коэффициент значим (это утверждение считается гарантированным при $\nu>20$ и $\alpha\ge 0{,}05$);
- если $3<|t|$, то коэффициент считается сильно значимым (вероятность ошибки при достаточном числе наблюдений не превосходит 0,001).

Каждая оценка дополняется доверительным интервалом. Для этого определяют предельную ошибку для каждого коэффициента:

$$\Delta_i = t_{\alpha/2,n-2}\cdot S_i, \qquad 4.9$$

откуда границы доверительных интервалов находятся по формуле:

$$b_i \pm \Delta_{bi}. \qquad 4.10$$

Коэффициент **детерминации** для парной регрессии совпадает с **квадратом коэффициента корреляции $R^2=r^2_{xy}$** и характеризует долю дисперсии **результативного** признака y, объясняемую регрессией в общей дисперсии результативного признака. Соответственно величина $1\text{-}R^2$ характеризует долю дисперсии y, вызванную влиянием неучтенных факторов в общей дисперсии признака y.

$$\underset{\text{общая сумма квадратов откл}}{\sum(y_i-\bar{y})^2} = \underset{\text{объясненная …регрессией}}{\sum(\hat{y}_i-\bar{y})^2} + \underset{\text{остаточная сумма}}{\sum(\hat{y}_i-y_i)^2} \qquad 4.11^*$$

Разделив обе части уравнения на общую сумму квадратов отклонений, получим:

$$1=\frac{\sum(\hat{y}_i-\bar{y})^2}{\sum(y_i-\bar{y})^2}+\frac{\sum(\hat{y}_i-y_i)^2}{\sum(y_i-\bar{y})^2},$$

$$1=R^2+\frac{\sum e_i^2}{\sum(y_i-\bar{y})^2}\Rightarrow R^2=1-\frac{\sum e_i^2}{\sum(y_i-\bar{y})^2}. \qquad 4.12$$

Таким образом, **коэффициент детерминации R^2** является мерой, позволяющей определить, в какой степени найденная прямая регрессии дает лучший результат для объяснения поведения зависимой переменной y, чем горизонтальная прямая $y=\bar{y}$. Очевидно, что **$0\le R^2\le 1$**. Откуда следует, что чем ближе он к единице, тем больше уравнение регрессии объясняет поведение фактических значений y. Поэтому хотелось бы стремятся построить регрессию с наибольшим значением R^2.

Корень квадратный из коэффициента детерминации называется **индексом корреляции** и обозначают **ρ_{xy}**.

Для проверки общего качества уравнения регрессии выдвигается предположение, что коэффициенты b_0 и b_1 **одновременно равны нулю**, тогда уравнение считают незначимым, в противном случае значимым. Данная гипотеза проверяется на основе **дисперсионного** анализа, при этом сравниваются объясненная и остаточная дисперсии:

H_0: $S^2_{\hat{y}}=S^2$ – уравнение незначимо,

H_1: $S^2_{\hat{y}}>S^2$ – уравнение значимо.

Строится F-статистика:

$$F=\frac{\sum(\hat{y}_i-\bar{y})^2/1}{\sum(y_i-\hat{y}_i)^2/(n-2)}=\frac{S^2_{\hat{y}}}{S^2}. \qquad 4.13$$

При выполнении условий МНК статистика имеет распределение Фишера с числом степеней свободы $\nu_1=1$, $\nu_2=n\text{-}1$. При уровне значимости α находят критичекую точку $F_{\alpha,1,n-1}=F_{кр}$ с помощью функции FИОБР и сравнивают его с наблюдаемым значением F. Так как рассматриваемая гипотеза правосторонняя [1], то:

- если $F>F_{кр}$, то гипотеза H_0 отклоняется в пользу H_1, что означает объясненная дисперсия существенно больше остаточной, следовательно, уравнение регрессии достаточно качественно отражает динамику изменения зависимой переменной от объясняющей.

- если $F<F_{кр}$, то гипотеза H_0 принимается, т.е. объясненная дисперсия соизмерима с остаточной дисперсией, вызванной случайными факторами. Это позволяет считать влияние объясняющих переменных модели несущественным, а следовательно, общее качество уравнения регрессии невысоким.

В случае **парной регрессии** проверка нулевой гипотезы для F-статистики равносильна проверке нулевой гипотезы для t_r-статистики для коэффициента корреляции:

$$t_r = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}},$$

Можно доказать равенство:

$$F = t_r^2 = t_1^2 \qquad 4.14$$

Самостоятельную значимость коэффициент R^2 приобретает в случае **множественной** регрессии.

ПОИСК ПРОГНОЗНОГО ЗНАЧЕНИЯ И ЕГО ОЦЕНКА

Прогнозное значение $\hat{y}_p$ определяется, если в уравнение регрессии подставить значение x_p:

$$\hat{y}_p = b_0 + b_1 x_p \ . \qquad 4.15$$

Границы доверительного интервала для параметра y_p будут равны:

$$\hat{y}_p \pm t_{\alpha/2,\, n-2} \cdot S_p. \qquad 4.16$$

Чтобы найти стандартную ошибку S_p прогнозного значения $\hat{y}_p$ можно использовать два подхода: либо рассматривать параметр y_p как отдельное значение переменной x_p; или разброс y_p найти как условное среднее значение при известном значении x_p.

Доверительный интервал для отдельного значения y_p учитывает источники рассеяния: для коэффициентов регрессии (1.5, 1.6) и всего уравнения регрессии (1.4). В этом случае **стандартная ошибка прогноза S_p** вычисляется по формуле:

$$S_p = S \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}, \qquad 4.17$$

Доверительный интервал для условного среднего не учитывает дисперсию для всего уравнения регрессии (1.4), поэтому формула для вычисления ошибки прогноза имеет вид:

$$S_{\hat{y}} = \sqrt{\frac{S^2}{n} + \frac{S^2 (x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \ . \qquad 4.18$$

Нелинейная парная регрессия

Многие экономические зависимости не являются линейными по своей сути, и поэтому их моделирование линейными уравнениями не даст положительного результата. Если между экономическими явлениями существуют нелинейные соотношения, то они выражаются с помощью нелинейных регрессий.

Все уравнения нелинейных регрессий можно разбить на два класса:

1. нелинейные относительно объясняющих переменных, но линейные по оцениваемым параметрам:

$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k + \varepsilon$ (полиномиальная – многочлен любой степени);

$y = \alpha + \dfrac{\beta}{x} + \varepsilon$ – равносторонняя гипербола;

$y = \alpha + \beta \ln x + \varepsilon$ – полулогарифмическая кривая и другие.

2. Нелинейные по оцениваемым параметрам.

$y = \alpha x^{\beta} \varepsilon$ – степенная;

$y = \alpha \beta^{x} \varepsilon$ – показательная;

$y = e^{\alpha + \beta x} \varepsilon$ –экспоненциальная.

Первый класс уравнений довольно просто сводится к линейной регрессии путем ввода новых факторов.

Нелинейная регрессия по включенным переменным не таит каких-либо сложностей в оценке ее параметров. Она определяся, как и в линейной регрессии, методом наименьших квадратов» (МНК), ибо эти функции линейны по параметрам. Так, в ***параболе второй степени*** (эмпирическая модель)

$\hat{y} = a_0 + a_1 x + a_2 x^2$,

заменяя переменные $x = x_l$, $x^2 = x_2$, получим двухфакторное уравнение линейной регрессии: $\hat{y} = a_0 + a_1 x_1 + a_2 \; x_2$, для определения параметров которого используется МНК.

Применение МНК для оценки параметров параболы второй степени приводит к следующей системе нормальных уравнений:

$$\sum y = n \cdot a_0 + a_1 \cdot \sum x + a_2 \cdot \sum x^2$$
$$\sum yx = a_0 \cdot \sum x + a_1 \cdot \sum x^2 + a_2 \cdot \sum x^3$$
$$\sum yx^2 = a_0 \cdot \sum x^2 + a_1 \cdot \sum x^3 + a_2 \cdot \sum x^4$$

Система содержит три уравнения с тремя неизвестными и еерешение может быть произведено по формулам Крамера.

Парабола второй степени $\hat{y} = a_0 + a_1 x + a_2 x^2$ при $a_1 > 0$ и $a_2 < 0$ симметрична относительно высшей точки, т. е. точки максимума, изменяющей направление связи, а именно — рост на падение. Такого рода функцию можно наблюдать в экономике труда при изучении зависимости заработной платы работников физического труда от возраста, с увеличением возраста повышается заработная плата ввиду одновременного увеличения опыта и повышения квалификации работника. Однако с определенного возраста ввиду старения организма и снижения производительности труда дальнейшее повышение возраста может приводить к снижению заработной платы работника.

При $a_1 < 0$ и $a_2 > 0$ парабола второго порядка симметрична относитело своей низшей точки, что позволяет определять минимум функции в точке, меняющей направление связи, т. е. снижение на рост.

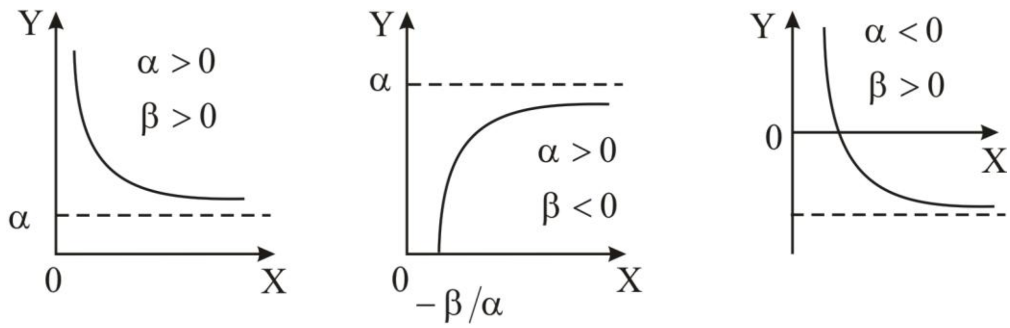
Ввиду симметричности кривой парабола второй степени далеко не всегда пригодна в конкретных исследованиях, чаще имеют дело лишь с отдельными сегментами параболы.

Среди класса нелинейных функций, параметры которых без особых затруднений оцениваются МНК, следует назвать хорошо известную в эконометрике ***равностороннюю гиперболу***:

$$Y = \alpha + \frac{\beta}{X} + \varepsilon$$

Она может быть использована на микро- и макроуровне, например, для характеристики связи удельных расходов сырья, материалов, топлива с объемом выпускаемой продукции, времени обращения товаров от величины товарооборота. Классическим ее примером является кривая Филлипса, английского экономиста, характеризующая соотношение между нормой безработицы и процентом прироста заработной платы.

В зависимости от знаков α и β возможны ситуации, изображенные на рис. 4.1



а *б* *в*

Рис. 4.1.

График на рис. 4.1. *а* может отражать зависимость между объемом выпуска (X) и средними фиксированными издержками (Y).

График на рис.4.1, *б* — зависимость между доходом X и спросом на блага Y (например, на товары первой необходимости либо товары относительной роскоши); это так называемые функции Торнквиста (в этом случае $X = -\dfrac{\beta}{\alpha}$ — минимально необходимый уровень дохода). Математическое описание подобного рода взаимосвязей получило название ***кривых Энгеля***. В 1857 г. немецкий статистик Э. Энгель на основе исследования семейных расходов сформулировал закономерность — с ростом дохода доля доходов, расходуемых на продовольствие, уменьшается. Соответственно с увеличением дохода доля доходов, расходуемых на непродовольственные товары, будет возрастать. Однако это увеличение не беспредельно, ибо на все товары сумма долей не может быть больше единицы.

Важным приложением графика, изображенного на рис. 4.1, *в*, является кривая Филлипса, отражающая зависимость между уровнем безработицы (X) в процентах и процентным изменением заработной платы (Y). При этом точка пересечения кривой с осью 0X определяет естественный уровень безработицы.

Для равносторонней гиперболы в уравнении регрессии заменив $\dfrac{1}{X}$ на Z, получим линейное уравнение регрессии $Y = \alpha + \beta Z + \varepsilon$, оценка параметров которого может быть дана МНК. Система нормальных уравнений составит:

$$\sum y = n \cdot a + b \cdot \sum \frac{1}{x}$$

$$\sum \frac{y}{x} = a \cdot \sum \frac{1}{x} + b \cdot \sum \frac{1}{x^2},$$

где a и b — оценки параметров регрессии.

При $b > 0$ имеем обратную зависимость, которая при $x \to \infty$ характеризуется нижней асимптотой, т. е. минимальным предельнIым значением *y,* оценкой которого.служит параметр *a.* При $b < 0$ имеем медленно повышающуюся функцию с верхней асимптотой при $x \to \infty$,.

Когда надо определить темп роста или прироста каких либо экономических показателей используют ***полулогарифмические кривые***

$$Y = \alpha + \beta \cdot \ln X + \varepsilon$$

$$\ln Y = \alpha + \beta \cdot X + \varepsilon$$

Рассмотрим зависимость, хорошо известную в банковском и финансовом анализе

$$Y_t = Y_0 (1 + r)^t,$$

где Y_0 – первоначальный вклад, r – сложный темп прироста, Y_t – вклад в момент времени t. Прологарифмируем уравнение (3.31) и введем обозначения: $\ln Y_0 = \alpha$, $\ln(1+r) = \beta$, получим модель вида (3.30): $\ln Y = \alpha + \beta \cdot t + \varepsilon_t$. Дополнительное случайное слагаемое ε_t характеризует возможность изменения процентной ставки.

Для оценки коэффициентов полулогарифмических кривых (3.30) используют МНК. Покажем применение МНК на примере кривой $Y = \alpha + \beta \cdot \ln X + \varepsilon$. Заменив *lnX* на *Z*, опять получим линейное уравнение: $Y = \alpha + \beta \cdot Z + \varepsilon$. Данная функция, как и предыдущая, линейна по параметрам и нелинейна по объясняющей переменной. Оценка параметров регрессии может быть найдена МНК. Система нормальных уравнений при этом окажется следующей:

$$\sum y = n \cdot a + b \sum \ln x$$
$$\sum y \cdot \ln x = a \cdot \sum \ln x + \sum (\ln x)^2,$$

где a и b — оценки параметров регрессии.

В моделях нелинейных по оцениваемым параметрам, но приводимых к линейному виду, МНК применяется к преобразованным уравнениям. Если в линейной модели и моделях, нелинейным по переменным, при оценке параметров исходят из критерия $\sum (y - \hat{y}_x)^2 \to \min$, то в моделях нелинейных по оцениваемым параметрам, требование МНК применяется не к исходным данным результативного признака, а к их преобразованным величинам, т.е. *lny, 1/y*. Вследствие этого оценка параметров для линеаризуемых функций МНК оказывается несколько смещенной.

Рассмотрим регрессию, нелинейную по оцениваемым параметрам (второго класса). Данный класс нелинейных моделей подразделяется на два типа: нелинейные модели внутренне линейные и нелинейные модели внутренне нелинейные. Если нелинейная модель внутренне линейна, то она с помощью соответствующих преобразований может быть приведена к линейному виду. Если же нелинейная модель внутренне нелинейная, то она не может быть сведена к линейной функции.

Например, в эконометрических исследованиях при изучении эластичности спроса от цен широко используется ***степенная регрессия***:

$$Y = \alpha \cdot X^{\beta} \cdot \varepsilon,$$

где α и β – параметры модели.

Степенная зависимость может отражать зависимость спроса на товар от его цены (тогда $\beta<0$) или дохода (в этом случае $\beta>0$). Функция (3.33) может так же характеризовать зависимость объема выпуска от количества используемого ресурса тогда $0<\beta<1$(производственная функция). Предполагается, что случайная ошибка связана с объясняющей переменной мультипликативно.

Параметр β имеет четкое экономическое истолкование: **показывает, на сколько процентов изменится в среднем результат, если фактор изменится на 1%,** т.е. он является **коэффициентом эластичности**.

Коэффициент эластичности вычисляется по формуле

$$Э = f'(x)\frac{x}{y},$$

где Э — коэффициент эластичности.

Для оценки параметров степенной регрессии применяется МНК к линеаризированному уравнению $\ln \hat{y} = \ln a + b \cdot \ln x$ ($\hat{y} = a \cdot x^b \cdot e$), т.е. решается система нормальных уравнений:

$$\sum \ln y = n \cdot \ln a + b \cdot \sum \ln x$$

$$\sum \ln y \cdot \ln x = \ln a \cdot \sum \ln x + b \cdot \sum (\ln x)^2,$$

где a и b — оценки параметров степенной регрессии.

Оценка параметра β степенной регрессии b определяется из системы, а оценка параметра α – a потенцированием выражения $\ln a$.

В случае парной регрессии обоснованность использования степенной модели проверить достаточно просто. Для этого на корреляционное поле наносятся точки (lnx_i, lny_i, i=1, 2,….n). Если их расположение соответствует прямой линии, то произведенная замена удачна и использование логарифмической модели обосновано.

Примером внутренне нелинейных моделей могут служить модели вида $Y = \alpha + \beta \cdot X^{\sigma} + \varepsilon$, $Y = \alpha[1 - 1/(1 - X^{\beta})] + \varepsilon$. Эти уравнения не могут быть преобразованы в уравнения, линейные по коэффициентам.

Если модель внутренне нелинейная по параметрам, то для оценки ее параметров используют итерационные методы. Такие модели могут иметь место в эконометрических исследованиях, однако, наибольшее распространение получили модели, приводимые к линейному виду. Решение такого типа моделей реализовано в стандартных пакетах прикладных программ.

Поскольку коэффициенты эластичности представляют экономический интерес, а виды моделей не ограничиваются только степенной функцией, приведем формулы расчета коэффициентов эластичности для наиболее распространенных типов уравнений регрессии (табл. 2.).

Таблица 2.

Вид функции	Коэффициент эластичности
Линейная $y = a + b \cdot x + \varepsilon$	$Э = \dfrac{b \cdot x}{a + b \cdot x}$
Парабола второго порядка $y = a + b \cdot x + c \cdot x^2 + \varepsilon$	$Э = \dfrac{(b + 2 \cdot c \cdot x) \cdot x}{a + b \cdot x + c \cdot x^2}$
Гипербола $y = a + \dfrac{b}{x} + \varepsilon$	$Э = \dfrac{-b}{b + a \cdot x}$
Показательная $y = \alpha \cdot \beta^x \cdot \varepsilon$	$Э = x \cdot \ln b$
Степенная $y = \alpha \cdot x^{\beta} \cdot \varepsilon$	$Э = b$
Полулогарифмическая $y = a + b \cdot \ln x + \varepsilon$	$Э = \dfrac{b}{a + b \cdot \ln x}$

В силу того, что коэффициент эластичности для многих функций не является величиной постоянной, а зависит от соответствующего значения x, то обычно рассчитывается средний показатель эластичности по формуле

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

Нелинейная корреляция

При отклонении парной статистической зависимости от линейной коэффициент корреляции теряет свой смысл как характеристика степени тесноты связи. В этом случае можно воспользоваться таким измерителем связи как индекс корреляции (корреляционное отношение).

$$\rho_{xy}{}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}.$$

Величина данного показателя находится в границах от 0 до 1. Нетрудно заметить, что для вычисления индекса корреляции используется та же формула, что и для коэффициента детерминации для линейной парной регрессии. Поэтому коэффициент корреляции имеет тот же смысл, что и коэффициент детерминации, т.е. $\rho_{xy}{}^2 = R^2$. В специальных исследованиях величину R^2 для нелинейных связей называют индексом детерминации.

Индекс детерминации используется для проверки существенности в целом уравнения нелинейной регрессии по F-критерию Фишера:

$$F_{набл} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2 \Big/ m}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2 \Big/ (n-m-1)} = \frac{\rho_{xy}{}^2}{1-\rho_{xy}{}^2} \cdot \frac{n-m-1}{m},$$

где n — число единиц совокупности, m —число параметров при переменных x. Величина m характеризует число степеней свободы для факторной суммы квадратов, а $(n-m-1)$ — число степеней свободы для остаточной суммы квадратов.

$F_{табл}$ — это максимально возможное значение критерия при m и $(n-m-1)$ степенях свободы и уровне значимости α.

Если $F_{табл} < F_{набл}$, то H_0 — гипотеза о случайной природе оцениваемых характеристик отклоняется и признается их статистическая незначимость, ненадежность уравнения регрессии .

Проверка качества уравнения регрессии

Третьим этапом решения задачи построения уравнения регрессии, соответствующего эмпирическим данным и целям исследования, является анализ качества уравнения.

Проверка качества уравнения регрессии заключается в статистической проверке значимости (надежности): уравнения регрессии, коэффициентов регрессии и корреляции, анализе остатков. Важность безусловного выполнения этих процедур обусловливается последующей практической направленностью полученных результатов.

Если определена функция регрессии и она экономически обоснована, а точность статистических оценок параметров соответствует предъявляемым требованиям, то прогнозируемые значения экономических показателей обладают достаточной надежностью.

Сущность процедур проверки значимости уравнения регрессии, коэффициентов регрессии и корреляции изложена выше. Напомним лишь те формулы, на основе которых осуществляется эта проверка.

1. Для оценки значимости коэффициентов регрессии используется t-критерий Стьюдента

2. Для оценки надёжности выборочного уравнения регрессий и проверки значимости коэффициента детерминации применяется F — критерий Фишера,

Чтобы иметь общее суждение о качестве модели из относительных отклонений по каждому наблюдению, определяют среднюю ошибку аппроксимации

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

.

Допустимый предел значений $\overline{A}$ не более 10-12%.

Регрессионный анализ. В регрессионном анализе моделируется взаимосвязь одной случайной переменной от одной или нескольких других случайных переменных. При этом, первая переменная называется зависимой, а остальные – независимыми. Выбор или назначение зависимой и независимых переменных является произвольным (условным) и осуществляется исследователем в зависимости от решаемой им задачи. Независимые переменные называются факторами, регрессорами или предикторами, а зависимая переменная – результативным признаком, или откликом.
Если число предикторов равно 1, регрессию называют простой, или однофакторной, если число предикторов больше 1 – множественной или многофакторной. В общем случае регрессионную модель можно записать следующим образом:

$$y = f(x1, x2, \ldots, xn),$$

где y – зависимая переменная (отклик),

xi (i = 1,…, n) – предикторы (факторы),

n – число предикторов.

Посредством регрессионного анализа можно решать ряд важных для исследуемой проблемы задач:

1). Уменьшение размерности пространства анализируемых переменных (факторного пространства), за счет замены части факторов одной переменной – откликом. Более полно такая задача решается факторным анализом.

2). Количественное измерение эффекта каждого фактора, т.е. множественная регрессия, позволяет исследователю задать вопрос (и, вероятно, получить ответ) о том, «что является лучшим предиктором для...». При этом, становится более ясным воздействие отдельных факторов на отклик, и исследователь лучше понимает природу изучаемого явления.

3). Вычисление прогнозных значений отклика при определенных значениях факторов, т.е. регрессионный анализ, создает базу для вычислительного эксперимента с целью получения ответов на вопросы типа «Что будет, если… ».

4). В регрессионном анализе в более явной форме выступает причинно-следственный механизм. Прогноз при этом лучше поддается содержательной интерпретации.

Раздел 5. Множественный регрессионный анализ.

В предыдущем пункте метод наименьших квадратов описан в простейшем случае. Он допускает различные обобщения. Например, метод наименьших квадратов дает алгоритм расчетов в случае, если исходные данные – по-прежнему набор n пар чисел (t_k, x_k), $k = 1,2,\ldots,n$, где t_k – независимая переменная (например, время), а x_k – зависимая (например, индекс инфляции - см. главу 7), а восстанавливать надо не линейную зависимость, а квадратическую:

$$x(t) = at^2 + bt + c.$$

Следует рассмотреть функцию трех переменных

$$f(a,b,c) = \sum_{k=1}^{n}(x_k - at_k^2 - bt_k - c)^2.$$

Оценки метода наименьших квадратов - это такие значения параметров a^*, b^* и c^*, при которых функция $f(a,b,c)$ достигает минимума по всем значениям аргументов. Чтобы найти эти оценки, надо вычислить частные производные от функции $f(a,b,c)$ по аргументам a, b и c, приравнять их 0, затем из полученных уравнений найти оценки: Имеем:

$$\frac{\partial f(a,b,c)}{\partial a}=\sum_{k=1}^{n}\frac{\partial}{\partial a}(x_k-at_k^2-bt_k-c)^2=\sum_{k=1}^{n}2(-t_k^2)(x_k-at_k^2-bt_k-c)^2.$$

Приравнивая частную производную к 0, получаем линейное уравнение относительно трех неизвестных параметров *a,b,c:*

$$a\sum_{k=1}^{n}t_k^4+b\sum_{k=1}^{n}t_k^3+c\sum_{k=1}^{n}t_k^2=\sum_{k=1}^{n}t_k^2x_k.$$

Приравнивая частную производную по параметру *b* к 0, аналогичным образом получаем уравнение

$$a\sum_{k=1}^{n}t_k^3+b\sum_{k=1}^{n}t_k^2+c\sum_{k=1}^{n}t_k=\sum_{k=1}^{n}t_kx_k.$$

Наконец, приравнивая частную производную по параметру *c* к 0, получаем уравнение

$$a\sum_{k=1}^{n}t_k^2+b\sum_{k=1}^{n}t_k+cn=\sum_{k=1}^{n}x_k.$$

Решая систему трех уравнений с тремя неизвестными, находим оценки метода наименьших квадратов.

Другие задачи, рассмотренные в предыдущем пункте (доверительные границы для параметров и прогностической функции и др.), также могут быть решены. Соответствующие алгоритмы более громоздки. Для их записи полезен аппарат матричной алгебры (см., например, одну из лучших в этой области монографий [2]). Для реальных расчетов используют соответствующие компьютерные программы.

Теория оценивания неизвестных параметров хорошо развита именно в случае линейного регрессионного анализа. Если же линейности нет и нельзя перейти к линейной задаче, то, как правило, хороших свойств от оценок ожидать не приходится.

Продемонстрируем подходы в случае зависимостей различного вида. Если зависимость имеет вид многочлена (полинома)

$$x(t)=a_0+a_1t+a_2t^2+a_3t^3+\ldots+a_mt^m,$$

то коэффициенты многочлена могут быть найдены путем минимизации функции

$$f(a_0,a_1,a_2,a_3,\ldots,a_m)=\sum_{k=1}^{n}(x_k-a_0-a_1t_k-a_2t_k^2-a_3t_k^3-\ldots-a_mt_k^m)^2.$$

Функция от *t* не обязательно должна быть многочленом. Можно, например, добавить периодическую составляющую, соответствующую сезонным колебаниям. Хорошо известно, например, что инфляция (рост потребительских цен) имеет четко выраженный годовой цикл - в среднем цены быстрее всего растут зимой, в декабре - январе, а медленнее всего (иногда в среднем даже падают) летом, в июле - августе. Пусть для определенности

$$x(t)=a_0+a_1t+a_2t^2+a_3t^3+\ldots+a_mt^m+A\sin Bt,$$

тогда неизвестные параметры могут быть найдены путем минимизации функции

$$f(a_0,a_1,a_2,a_3,\ldots,a_m,A,B)=\sum_{k=1}^{n}(x_k-a_0-a_1t_k-a_2t_k^2-a_3t_k^3-\ldots-a_mt_k^m-A\sin Bt_k)^2.$$

Пусть $I(t)$ -индекс инфляции в момент t. Принцип стабильности условий приводит к гипотезе о постоянстве темпов роста средних цен, т.е. индекса инфляции. Таким образом, естественная модель для индекса инфляции - это

$$I(t)=Ae^{Bt}.$$

Эта модель не является линейной, метод наименьших квадратов непосредственно применять нельзя. Однако если прологарифмировать обе части предыдущего равенства:

$$\ln I(t)=\ln A+Bt,$$

то получим линейную зависимость, рассмотренную в первом пункте настоящей главы.

Независимых переменных может быть не одна, а несколько. Пусть, например, по исходным данным $(x_k, y_k, z_k), k = 1,2,\dots,n,$ требуется оценить неизвестные параметры a и b в зависимости

$$z = ax + by + \varepsilon,$$

где ε - погрешность. Это можно сделать, минимизировав функцию

$$f(a,b) = \sum_{k=1}^{n} (z_k - ax_k - by_k)^2.$$

Зависимость от x и y не обязательно должна быть линейной. Предположим, что из каких-то соображений известно, что зависимость должна иметь вид

$$z = ax + by + cx^2 y + dxy + ey^3 + \varepsilon,$$

тогда для оценки пяти параметров необходимо минимизировать функцию

$$f(a,b,c,d,e) = \sum_{k=1}^{n} (z_k - ax_k - by_k - cx_k^2 y_k - dx_k y_k - ey_k^3)^2.$$

Более подробно рассмотрим пример из микроэкономики. В одной из оптимизационных моделей поведения фирмы используется т.н. производственная функция *f(K,L)*, задающая объем выпуска в зависимости от затрат капитала K и труда L. В качестве конкретного вида производственной функции часто используется так называемая функция Кобба-Дугласа

$$f(K,L) = K^{\alpha} L^{\beta}.$$

Однако откуда взять значения параметров α и β? Естественно предположить, что они - одни и те же для предприятий отрасли. Поэтому целесообразно собрать информацию $(f_k, K_k, L_k), k = 1,2,\dots n$ где f_k - объем выпуска на k-ом предприятии, K_k- объем затрат капитала на k-ом предприятии, L_k - объем затрат труда на k-ом предприятии (в кратком изложении здесь не пытаемся дать точных определений используемым понятиям из экономики предприятия). По собранной информации естественно попытаться оценить параметры α и β. Но они входят в зависимость нелинейно, поэтому сразу применить метод наименьших квадратов нельзя. Помогает логарифмирование:

$$\ln f(K,L) = \alpha \ln K + \beta \ln L.$$

Следовательно, целесообразно сделать *замену переменных*

$$x_k = \ln K_k, y_k = \ln L_k, z_k = \ln f_k, k = 1,2,3,\dots,n,$$

а затем находить оценки параметров α и β, минимизируя функцию

$$g(\alpha, \beta) = \sum_{k=1}^{n} (z_k - \alpha x_k - \beta y_k)^2.$$

Найдем частные производные:

$$\frac{\partial g(\alpha, \beta)}{\partial \alpha} = \sum_{k=1}^{n} 2(z_k - \alpha x_k - \beta y_k)(-x_k),$$

$$\frac{\partial g(\alpha, \beta)}{\partial \beta} = \sum_{k=1}^{n} 2(z_k - \alpha x_k - \beta y_k)(-y_k).$$

Приравняем частные производные к 0, сократим на 2, раскроем скобки, перенесем свободные члены вправо. Получим систему двух линейных уравнений с двумя неизвестными:

$$\alpha \sum_{k=1}^{n} x_k y_k + \beta \sum_{k=1}^{n} y_n^2 = \sum_{k=}^{n}$$

$$\alpha\sum_{k=1}^{n} x_k^2 + \beta\sum_{k=1}^{n} x_k y_k = \sum_{k=1}^{n} x_k z_k,$$

Таким образом, для вычисления оценок метода наименьших квадратов необходимо найти пять сумм

$$\sum_{k=1}^{n} x_k^2, \quad \sum_{k=1}^{n} x_k y_k, \quad \sum_{k=1}^{n} y_k^2, \quad \sum_{k=1}^{n} x_k z_k, \quad \sum_{k=1}^{n} y_k z_k.$$

Для упорядочения расчета этих сумм может быть использована таблица типа той, что применялась в первом пункте настоящей главы. Отметим, что рассмотренная там постановка переходит в разбираемую сейчас при $y_k = 1, \quad k = 1,2,\ldots,n.$

Основной показатель качества регрессионной модели. Одни и те же данные можно обрабатывать различными способами. Показателем отклонений данных от модели служит остаточная сумма квадратов *SS*. Чем этот показатель меньше, тем приближение лучше, значит, и модель лучше описывает реальные данные. Однако это рассуждение годится только для моделей с одинаковым числом параметров. Ведь если добавляется новый параметр, по которому можно минимизировать, то и минимум, как правило, оказывается меньше.

В качестве основного показателя качества регрессионной модели используют оценку остаточной дисперсии

$$\bar{\sigma}^2(m) = \frac{SS}{n-m},$$

скорректированную на число m параметров, оцениваемых по наблюдаемым данным. В случае линейной прогностической модели, рассмотренной в первом пункте настоящей главы, оценка остаточной дисперсии имеет вид

$$\bar{\sigma}^2 = \frac{SS}{n-2},$$

поскольку число оцениваемых параметров m=2.

Почему эта формула отличается от приведенной в первом пункте? Там в знаменателе n, а здесь - (n-2). Дело в том, что в первом пункте рассмотрена непараметрическая теория при большом объеме данных (при $n \to \infty$) , а при безграничном возрастании n разница между n и (n-2) сходит на нет.

А вот при подборе вида модели знаменатель дроби, оценивающей остаточную дисперсию, приходится корректировать на число параметров. Если этого не делать, то придется заключить, что многочлен второй степени лучше соответствует данным, чем линейная функция, многочлен третьей степени лучше приближает исходные данные, чем многочлен второй степени, и т.д. В конце концов доходим до многочлена степени (n-1) с n коэффициентами, который проходит через все заданные точки. Но его прогностические возможности, скорее всего, существенно меньше, чем у линейной функции. *Излишнее усложнение эконометрических моделей вредно.*

Типовое поведение скорректированной оценки остаточной дисперсии

$$v(m) = \bar{\sigma}^2(m)$$

в зависимости от параметра m в случае расширяющейся системы эконометрических моделей выглядит так. Сначала наблюдаем заметное убывание. Затем оценка остаточной дисперсии колеблется около некоторой константы (теоретического значения дисперсии погрешности).

Поясним ситуацию на примере эконометрической модели в виде многочлена

$$x(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + \ldots + a_m t^m.$$

Пусть эта модель справедлива при $m = m_0$. При $m < m_0$ в скорректированной оценке остаточной дисперсии учитываются не только погрешности измерений, но и соответствующие (старшие) члены многочлена (предполагаем, что коэффициенты при них отличны от 0). При $m \geq m_0$ имеем

$$\lim_{n\to\infty} v(m) = \sigma^2.$$

Следовательно, скорректированная оценка остаточной дисперсии будет колебаться около указанного предела. Поэтому в качестве оценки неизвестной эконометрику степени многочлена (полинома) можно использовать первый локальный минимум скорректированной оценки остаточной дисперсии, т.е.

$$m^* = \min\{\, m: v(m-1) > v(m), \quad v(m) \leq v(m+1)\}.$$

В работе [3] найдено предельное распределение этой оценки степени многочлена.

Теорема. При справедливости некоторых условий регулярности

$$\lim_{n\to\infty} P(m^* < m_0) = 0, \quad \lim_{n\to\infty} P(m^* = m_0 + u) = \lambda(1-\lambda)^u, \quad u = 0,1,2,\ldots,$$

где

$$\lambda = \Phi(1) - \Phi(-1) = \frac{1}{\sqrt{2\pi}} \int_{-1}^{1} \exp\left\{-\frac{x^2}{2}\right\} dx \approx 0{,}68268.$$

Таким образом, предельное распределение оценки m^* степени многочлена (полинома) является геометрическим. Это означает, в частности, что оценка не является состоятельной. При этом вероятность получить меньшее значение, чем истинное, исчезающе мала. Далее имеем:

$$P(m^* = m_0) \to 0{,}68268, \quad P(m^* = m_0 + 1) \to 0{,}68268(1 - 0{,}68268) = 0{,}21663,$$

$$P(m^* = m_0 + 2) \to 0{,}68268(1 - 0{,}68268)^2 = 0{,}068744,$$

$$P(m^* = m_0 + 3) \to 0{,}68268(1 - 0{,}68268)^3 = 0{,}021814\ldots$$

Разработаны и иные методы оценивания неизвестной степени многочлена, например, с помощью многократного применения процедуры проверки адекватности регрессионной зависимости с помощью статистики Фишера (см. работу [3]). Предельное поведение оценок - таково же, как в приведенной выше теореме, только значение параметра λ иное.

Линейный и непараметрические парные коэффициенты корреляции. Термин "корреляция" означает "связь". В эконометрике этот термин обычно используется в сочетании "коэффициенты корреляции".

Рассмотрим способы измерения связи между двумя случайными переменными. Пусть исходными данными является набор случайных векторов $(x_i, y_i) = (x_i(\omega), y_i(\omega)),\ i = 1,2,\ldots n.$ Коэффициентом корреляции, более подробно, линейным парным коэффициентом корреляции К. Пирсона называется (см. приложение 1 в конце настоящей книги)

$$r_n = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Если $r_n = 1$, то $y_i = ax_i + b,$ причем $a>0$. Если же $r_n = -1$, то $y_i = ax_i + b,$ причем $a<0$. Таким образом, близость коэффициента корреляции к 1 (по абсолютной величине) говорит о достаточно тесной линейной связи.

Коэффициенты корреляции типа r_n используются во многих алгоритмах многомерного статистического анализа эконометрических данных. В теоретических рассмотрениях часто считают, что случайный вектор имеет многомерное нормальное распределение. Распределения реальных данных, как правило, отличны от нормальных (см. главу 4). Почему же распространено представление о многомерном нормальном распределении? Дело в том, что теория в этом случае проще. В частности, равенство 0 теоретического коэффициента корреляции (см. приложение 1) эквивалентно независимости случайных величин. Поэтому проверка независимости сводится к проверке статистической гипотезы о равенстве 0 теоретического коэффициента корреляции. Эта гипотеза принимается, если $|r_n| < C(n,\alpha)$, где $C(n,\alpha)$ - некоторое граничное значение, зависящее от объема выборки n и уровня значимости α.

Если случайные вектора $(x_i, y_i) = (x_i(\omega), y_i(\omega)),\ i = 1,2,...n,$ независимы и одинаково распределены, то выборочный коэффициент корреляции сходится к теоретическому при безграничном возрастании объема выборки:

$$r_n \to \rho = \frac{M(x_1 - M(x_1))(y_1 - M(y_1))}{\sqrt{D(x_1)}\sqrt{D(y_1)}}$$

(сходимость по вероятности).

Более того, выборочный коэффициент корреляции является асимптотически нормальным. Это означает, что

$$\lim_{n\to\infty} P\left(\frac{r_n - \rho}{\sqrt{D_0(r_n)}} < x\right) = \Phi(x),$$

где $\Phi(x)$ - функция стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1, а $D_0(r_n)$ - асимптотическая дисперсия выборочного коэффициента корреляции. Она имеет довольно сложное выражение, приведенное в монографии [4, с.393]:

$$D_0(r_n) = \frac{\rho^2}{4n}\left(\frac{\mu_{40}}{\mu_{20}^2} + \frac{\mu_{04}}{\mu_{02}^2} + \frac{2\mu_{22}}{\mu_{20}\mu_{02}} + \frac{4\mu_{22}}{\mu_{11}^2} - \frac{4\mu_{31}}{\mu_{11}\mu_{20}} - \frac{4\mu_{13}}{\mu_{11}\mu_{02}}\right).$$

Здесь под μ_{km} понимаются теоретические центральные моменты порядка k и m, а именно,

$$\mu_{km} = M(x_1 - M(x_1))^k (y_1 - M(y_1))^m$$

(см. приложение 1 в конце книги).

Для расчета непараметрического *коэффициента ранговой корреляции Спирмена* необходимо сделать следующее. Для каждого x_i рассчитать его ранг r_i в вариационном ряду, построенном по выборке $x_1, x_2, ... x_n$. Для каждого y_i рассчитать его ранг q_i в вариационном ряду, построенном по выборке $y_1, y_2, ... y_n$. Для набора из n пар $(r_i, q_i),\ i = 1,2,...,n,$ вычислить (линейный) коэффициент корреляции. Он называется коэффициентом ранговой корреляции, поскольку определяется через ранги. В качестве примера рассмотрим данные из табл.2 (см. монографию [5]).

Табл.2. Данные для расчета коэффициентов корреляции

i	1	2	3	4	5
x_i	5	10	15	20	25
y_i	6	7	30	81	300
r_i	1	2	3	4	5
q_i	1	2	3	4	5

Для данных табл.2 коэффициент линейной корреляции равен 0,83, непосредственной линейной связи нет. А вот коэффициент ранговой корреляции равен 1, поскольку увеличение одной переменной однозначно соответствует увеличению другой переменной. Во многих экономических задачах, например, при выборе инвестиционных проектов для осуществления, достаточно именно монотонной зависимости одной переменной от другой.

Поскольку суммы рангов и их квадратов нетрудно подсчитать, то *коэффициент ранговой корреляции Спирмена* равен

$$\rho_n = 1 - \frac{6\sum_{i=1}^{n}(r_i - q_i)^2}{n^3 - n}.$$

Отметим, что *коэффициент ранговой корреляции Спирмена* остается постоянным при любом строго возрастающем преобразовании шкалы измерения результатов наблюдений. Другими словами, он является адекватным в порядковой шкале, как и другие ранговые статистики (см. статистики Вилкоксона, Смирнова), типа омега-квадрат для проверки однородности независимых выборок.

Широко используется также коэффициент ранговой корреляции τ Кендалла, коэффициент ранговой конкордации Кендалла и Б. Смита и др. Необходимые для практических расчетов таблицы имеются в справочник]. Дискуссия о выборе вида коэффициентов корреляции продолжается до настоящего времени.

СРАВНЕНИЕ КОЭФФИЦИЕНТОВ РЕГРЕССИИ

Простое сопоставление коэффициентов регрессии по модулю не может оценить силу влияния факторов на признак y: такое сопоставление лишено смысла. Однако их можно **нормировать** (**стандартизировать**), используя формулу:

$$\alpha_j = b_j \frac{S_{xj}}{S_y},$$

где α_j- коэффициент регрессии после нормирования,

S_j – стандартная ошибка переменной x_j,

S_y – стандартная ошибка переменной y.

Нормированные коэффициенты можно сравнивать и делать вывод о влиянии факторов на переменную y. Факторы с наименьшим по модулю значением α_j оказывают на y наименьшее влияние.

Уравнение регрессии в стандартизованном масштабе имеет вид:

$$t_y = \alpha_1 X1 + \alpha_2 X2 + ... + \alpha_m Xm$$

КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ

Коэффициент детерминации находится по формуле:

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

Он характеризует долю разброса значений зависимой переменной Y, объясненной уравнением регрессии.

СКОРРЕКТИРОВАННЫЙ КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ

В случае множественной регрессии коэффициент детерминации является **неубывающей** функцией числа **объясняющих переменных**, т.е. добавление новой переменной **увеличивает значение R^2**. Поэтому при расчете коэффициента детерминации для получения несмещенных оценок в числителе и знаменателе делается поправка на число степеней свободы. Найденное значение называется **скорректированным коэффициентом детерминации:**

$$\bar{R}^2 = 1 - \frac{\sum e_i^2 / (n-m-1)}{\sum (y_i - \bar{y})^2 / (n-1)}$$

где

- $\sum e_i^2 / (n-m-1) = S^2$ - является несмещенной **оценкой остаточной дисперсии,** т.е. дисперсией случайных отклонений точек наблюдений от линии регрессии. Ее число степеней свободы равно n-m-1, где m+1 степень свободы связана с необходимостью решения системы m+1 линейного уравнения;
- $\sum (y_i - \bar{y})^2 / (n-1) = S_{\bar{y}}^2$- является несмещенной **оценкой общей дисперсии**, т.е. дисперсией отклонения Y от $\bar{Y}$, где одна степень теряется при вычислении $\bar{Y}$.

Заметим, что несмещенная оценка **объясненной дисперсии** ($S_{\hat{y}}^2$), т.е. дисперсии отклонения точек $\hat{Y}$ от $\bar{y}$, имеет m степеней свободы.

Все суммы можно найти в столбце SS дисперсионного анализа, их средние значения в столбце MS, а число степеней свободы в столбце df этого же блока (Excel).

Можно получить формулу, устанавливающую связь между скорректированным коэффициентом детерминации и коэффициентом детерминации:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-m-1}.$$

Очевидно, что:

$\bar{R}^2 < R^2$ для m >1,

$\bar{R}^2 = R^2$ только при m=1.

$\bar{R}^2$ может принимать отрицательные значения (например, если R^2=0).

Коэффициент корректируется с ростом числа объясняющих переменных.

Доказано, что скорректированный коэффициент корреляции увеличивается при добавлении новой переменной тогда и только тогда, когда t- статистика этой переменной по модулю больше единицы. Поэтому **добавление** в модель новых переменных осуществляется до тех пор, пока он растет.

В пакете Анализ данных приводятся значения $\bar{R}^2$ и R^2. Значимость коэффициента детерминации и скорректированного коэффициента при исследовании уравнения регрессии большая, однако, не абсолютная. При **неправильной спецификации** модели можно получить очень высокие значения этих коэффициентов, поэтому $\bar{R}^2$ и R^2 рассматриваются как один из ряда показателей, которые нужно проанализировать, чтобы уточнить строящуюся модель.

ИНДЕКС МНОЖЕСТВЕННОЙ КОРРЕЛЯЦИИ

Теснота линейной взаимосвязи в линейной регрессии выполняется с помощью индекса корреляции:

$$\rho = \sqrt{1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2}}.$$

Если X – неслучайная величина, то ρ характеризует качество подбора уравнения регрессии. Если же X – случайная переменная, то индекс корреляции является мерой тесноты линейной взаимосвязи между Y и набором факторов X.

КОЭФФИЦИЕНТЫ ЧАСТНОЙ КОРРЕЛЯЦИИ.

Используются для выделения **определяющего** фактора и **второстепенных.** Необходимо определить частные зависимости между Y и X_j, при условии, что воздействие остальных факторов исключено (элиминировано). В случае трех переменных x, y, z. можно получить коэффициенты парной корреляции r_{xy}, r_{xz} и r_{yz} по формулам:

$$r_{yx\cdot z} = \frac{r_{yx} - r_{yz}\cdot r_{xz}}{\sqrt{1-r_{xz}^2}\cdot\sqrt{1-r_{yz}^2}} \quad r_{yz\cdot x} = \frac{r_{yz} - r_{yx}\cdot r_{zx}}{\sqrt{1-r_{zx}^2}\cdot\sqrt{1-r_{yx}^2}} \quad r_{zx\cdot y} = \frac{r_{zx} - r_{yz}\cdot r_{xy}}{\sqrt{1-r_{xy}^2}\cdot\sqrt{1-r_{yz}^2}}.$$

ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ ПРОГНОЗА

Если уравнение регрессии имеет вид: $Y= b_0+b_1X_1+b_2X_2+\ldots\ldots b_mX_m$,
то прогнозное значение вычисляется так же как в случае парной регрессии. Необходимо подставить заданные значения прогноза $X_p=(1, x_{p1}, x_{p2}, \ldots..x_{pm})$ в уравнение регрессии.

АНАЛИЗ КАЧЕСТВА УРАВНЕНИЯ РЕГРЕССИИ

Первое построенное по выборке уравнение редко является удовлетворительным по тем или иным характеристикам. Поэтому следующей задачей эконометрического анализа является проверка качества уравнения регрессии. Эта проверка проводится по следующим этапам:

- проверка статистической значимости коэффициентов регрессии;
- проверка общего качества уравнения регрессии;
- проверка свойств данных: проверка выполнимости МНК.

АНАЛИЗ ЗНАЧИМОСТИ R^2

Проверяется гипотеза об одновременном равенстве нулю всех объясняющих переменных – уравнение считается незначимым:

$$H_o: b_1= b_2=\ldots b_m=0.$$

Если данная гипотеза не отклоняется, то делается вывод, что совокупное влияние всех m объясняющих переменных на зависимую переменную Y можно считать статистически незначимым, а общее качество уравнения регрессии невысоким.

Проверка данной гипотезы проводится на основе дисперсионного анализа, при этом сравниваются объясненная и остаточная дисперсии.

$$H_0:\ S_{\hat{y}}^2=S^2$$

$$H_1:\ S_{\hat{y}}^2>S^2$$

Для проверки гипотезы строится F-статистика:

$$F=\frac{\sum(\hat{y}_i-\bar{y})^2/m}{\sum(y_i-\hat{y}_i)^2/(n-m-1)}=\frac{S_{\hat{y}}^2}{S^2},$$

которая при выполнении МНК имеет распределение Фишера с числом степеней свободы $\nu_1=m$, $\nu_2=n-m-1$. Критическое значение находится с помощью:

FИОБР() = $F_{\alpha,\ m,.n-m-1} = F_{кр}$ при уровне значимости α.

- Если $F>F_{кр}$ то гипотеза H_0 отклоняется в пользу H_1, что означает объясненная дисперсия существенно больше остаточной, следовательно, уравнение регрессии достаточно качественно отражает динамику изменения зависимой переменной от объясняющей.
- Если $F<F_{кр}$, то гипотеза H_0 принимается, т.е. объясненная дисперсия соизмерима с остаточной дисперсией, вызванной случайными факторами. Это позволяет считать влияние объясняющих переменных модели несущественным, а следовательно, общее качество уравнения регрессии невысоким.

На практике вместо указанной гипотезы проверяется, связанная с ней гипотеза о статистической значимости коэффициента детерминации R^2.

$$H_0: R^2=0$$

$$H_1: R^2>0$$

Очевидно, что если F=0, то и R^2 =0, а линия регрессии $Y= \bar{y}$ является наилучшей по МНК, т.е. величина Y\` линейно не зависит от X_1, X_2, $\ldots.X_m$. Анализ статистики F позволяет сделать вывод о том, что для принятия гипотезы об одновременном равенстве нулю всех коэффициентов линейной регрессии коэффициент детерминации R^2 не должен существенно отличаться от нуля. Его критическое значение уменьшается при росте числа наблюдений и может стать сколь угодно малым.

Для проверки этой гипотезы числитель и знаменатель формулы 1.29 поделим на общую сумму квадратов отклонений $\sum(y_i-\bar{y})^2$ и получим:

$$F=\frac{\sum(\hat{y}_i-\bar{y})^2/\sum(y_i-\bar{y})^2}{\sum(y_i-\hat{y}_i)^2/\sum(y_i-\bar{y})^2}\cdot\frac{n-m-1}{m}=\frac{R^2}{1-R^2}\cdot\frac{n-m-1}{m}.$$

Вернемся к результатам нашего примера 1.3. (рис.1.14).Найдем по таблице распределения Фишера критическую точку для уровня значимости α: $F_{0,05;2;8}=4,46$. Сравнивая критическое и наблюдаемое значения (F=175,47), можно сделать вывод, что коэффициент детерминации статистически значим. Это означает, что совокупное влияние переменных X и Y на переменную C существенно. Этот же вывод можно сделать по столбцу значимость F, который характеризует вероятность выполнения гипотезы H_0.

ПРОВЕРКА КАЧЕСТВА ДВУХ КОЭФФИЦИЕНТОВ ДЕТЕРМИНАЦИИ

Статистику F можно использовать и для обоснования случая **исключения** или **добавления** в уравнение регрессии k объясняющих переменных.

Добавлять (исключать) переменные надо по одному.

Использовать лучше $\bar{R}^2$, так как R^2 всегда растет при добавлении новой объясняющей переменной.

Зависимая переменная должна быть представлена в том же виде, что и уже существующие в исследуемом уравнении регрессии. Например, Y моделируется двумя уравнениями: линейным

Число наблюдений для обеих моделей должно быть одинаковым.

Пусть первоначально построенное по n наблюдениям уравнение регрессии имело вид:

$$\hat{Y}=b_0+b_1X_1+b_2X_2+.....+b_{m-k}X_{m-k}+....+b_mX_m,$$

и скорректированный коэффициент детерминации равен $\bar{R}_1^2$.

Исключим из уравнения k переменных, оказывающих наименьшее влияние на Y. По n наблюдениям построим новое уравнение регрессии:

$$Y=c_0+c_1X_1+c_2X_2+.....+c_{m-k}X_{m-k},$$

скорректированный коэффициент детерминации, для которого равен $\bar{R}_2^2$.

Необходимо определить существенно ли ухудшилось качество описания зависимой переменной Y. Для этого выдвинем гипотезы:

H_0: $\bar{R}_1^2-\bar{R}_2^2=0$ – ничего не изменилось

H_1: $\bar{R}_1^2-\bar{R}_2^2\neq 0$ – уравнение ухудшилось, если разность больше нуля.

По выборочным данным найдите статистику:

$$F=\frac{\bar{R}_1^2-\bar{R}_2^2}{1-\bar{R}_1^2}\cdot\frac{n-m-1}{k},$$

которая имеет распределения Фишера с числом степеней свободы $\nu_1=k$, $\nu_2=n-m-1$, где

$(\bar{R}_1^2-\bar{R}_2^2)/k$ – потеря качества уравнения в результате того, что k переменных было отброшено. В результате появляется k дополнительных степеней свободы;

$(1-\bar{R}_1^2)/(n-m-1)$ – остаточная дисперсия первоначального уравнения.

Сравним критическое значение $F_{кр}$ и с наблюдаемым при уровне значимости α:

- Если $|F|>F_{kp}$, то гипотеза H_0 отклоняется в пользу H_1, что означает, одновременное исключение k объясняющих переменных существенно повлияет на качество первоначального уравнения.
- Если $|F|<F_{kp}$, то гипотеза H_0 принимается, т.е. разность $\bar{R}_1^2-\bar{R}_2^2$ незначительная. Это позволяет считать, что исключение k объясняющих переменных модели допустимым, так как общее качество уравнения регрессии изменится несущественно.

Аналогично проверяется гипотеза о добавлении k объясняющих переменных в уравнение регрессии. В этом случае составляется статистика:

$$F = \frac{\bar{R}_2^2 - \bar{R}_1^2}{1 - \bar{R}_2^2} \cdot \frac{n - m - 1}{k}.$$

ПРОВЕРКА ВЫПОЛНИМОСТИ МНК. АВТОКОРРЕЛЯЦИЯ ОСТАТКОВ. СТАТИСТИКА ДАРБИНА–УОТСОНА.

Все предыдущие рассуждения основаны на том, что выполняются предпосылки МНК: мы предполагали, что случайные отклонения e_i являются независимыми случайными величинами со средней, равной нулю. При работе с фактическими данными, такое допущение не всегда выполняется. Например, если вид функции выбран неудачно, то отклонения от регрессии вряд ли будут независимыми. В этом случае замечается концентрация положительных или отрицательных отклонений от регрессии и можно сомневаться в их случайном характере.

Если последовательные значения e_i коррелируют (зависят) между собой, то говорят, что имеет место **автокорреляция остатков**.

МНК в случае автокорреляции дает несмещенные и состоятельные оценки, однако полученные в этом случае доверительные интервалы имеют мало смысла в силу своей ненадежности. Значительная автокорреляция говорит о том, что спецификация модели неправильная. Проверка остатков на автокорреляцию должна выполняться обязательно. Наиболее простым приемом обнаружения автокорреляции является метод Дарбина–Уотсона (DW). Идея, которого состоит в том, что проверяются на коррелированность не любые, а только соседние величины e_i. Соседними обычно считаются соседние по возрастанию объясняющей переменной X (в случае перекрестной выборки) или по времени (в случае временных рядов) значения e_i.

Статистика DW рассчитывается по формуле:

$$d = DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}.$$

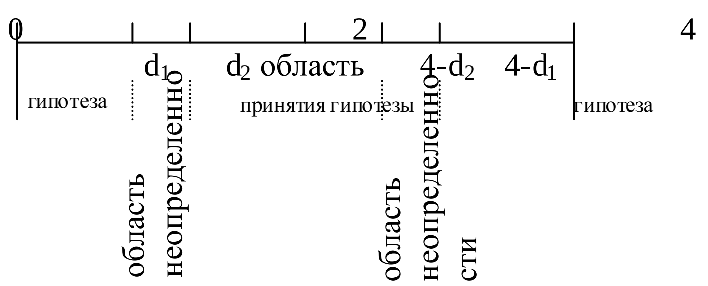
При условии что $M(e_i)=0$, $(i=1,2,\ldots.n)$ и n большое число можно предположить $\sum_{i=1}^{n} e_i^2 \approx \sum_{i=2}^{n} e_{i-1}^2$, тогда после преобразования получим:

$$d = \frac{2(\sum_{i=2}^{n} e_i^2 - \sum_{i=2}^{n} e_i e_{i-1})}{\sum_{i=1}^{n} e_i^2} = 2 - 2\frac{\sum_{i=2}^{n} e_i e_{i-1}}{\sum_{i=1}^{n} e_i^2}.$$

Очевидно, что $0 \le d \le 4$, так как коэффициент корреляции $\left| r_{e_i e_{i-1}} \right| \le 1$

- $d=2$, если $cov(e_i, e_{i-1})=0$ – автокорреляция отсутствует;
- $d=0$ –полная положительная автокорреляция;
- $d=4$ –полная отрицательная автокорреляция.

Возникает вопрос, какие значения d можно считать близкими к 2? Для обнаружения границ наблюдений статистики d существуют специальные таблицы. Для заданных α-уровня значимости; n- числа наблюдений и m –числа объясняющих переменных указывается два числа: d_1- нижняя граница и d_2- верхняя граница. Не обращаясь к таблице критических точек DW можно воспользоваться правилом, если 1,5<d<2,5, автокорреляция отсутствует. Изобразим на рисунке числовой отрезок, используемый для проверки гипотезы об отсутствии автокорреляции.



отклоняется отклоняется

положительная автокорреляция

отрицательная автокорреляция

Рис. 1. Область для принятия гипотезы об отсутствии автокорреляции.

Для вычисления этой статистики запустите инструмент Регрессия, включив опции **Остатки** и **График остатков**, как показано на рис. 2. В результате получите значение случайных отклонений e_i и их графики, которые Excel строит для каждой независимой переменной, как показано на рис.1.20 и 1.21. Чтобы найти DW, можно использовать функции СУММКВРАЗН и СУММКВ.

Если зависимость между С и X линейная, то график остатков должен иметь случайный вид. На рис.4 видим систематический рисунок, поэтому скорее всего между С и Y существует нелинейная зависимость, а значит надо изменить модель, включая в нее нелинейную зависимость.

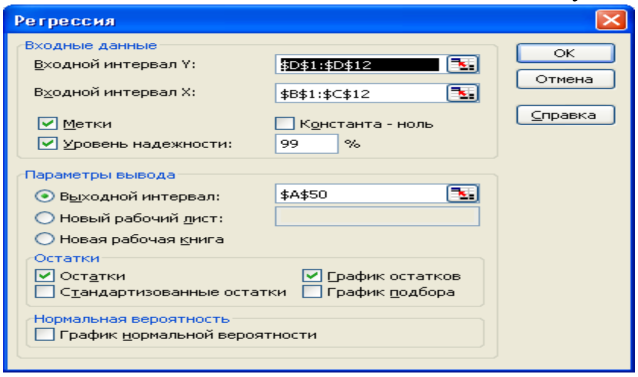


Рис. 2. Окно Регрессии для изучения автокорреляции остатков

Для проверки статистической значимости DW надо воспользоваться таблицей критических точек Дарбина-Уотсона, например, при уровне значимости α=0,05 и числе наблюдений n=11 d_1=0,658; d_2=1,604. Можно считать, что автокорреляция отсутствует, так как найденная статистика попадает в критический интервал: 1,604<DW<2,396, что является подтверждением высокого качества модели.

	A	B	C	D	E
72	ВЫВОД ОСТАТКА				
73					
74	аблюден	редсказанное	Остатки	ei-e(i-1)	(ei-e(i-1))2
75	1	22,48852	-2,48852		
76	2	23,73040854	1,269591	3,758111458	14,123402
77	3	31,009917	-1,00992	-2,279508462	5,1961588
78	4	28,6979627	1,302037	2,311954296	5,3451327
79	5	33,49369408	1,506306	0,204268621	0,0417257
80	6	37,04753692	0,952463	-0,553842837	0,3067419
81	7	39,531314	0,468686	-0,483777083	0,2340403
82	8	38,46124825	-0,46125	-0,929934246	0,8647777
83	9	45,74075671	-1,74076	-1,279508462	1,6371419
84	10	51,77837663	-1,77838	-0,03761992	0,0014153
85	11	53,02026517	1,979735	3,758111458	14,123402
86				Сумма	41,873938

Рис. 3. Вычисление наблюдаемой статистики d по формуле 1.33

X График остатков

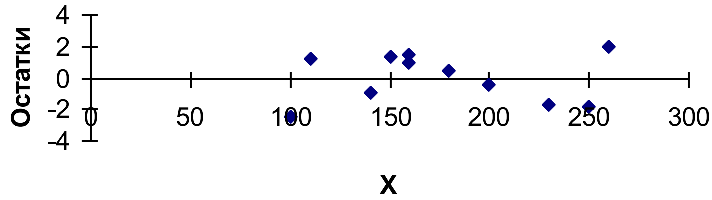


Рис. 4. График остатков регрессии

Y График остатков

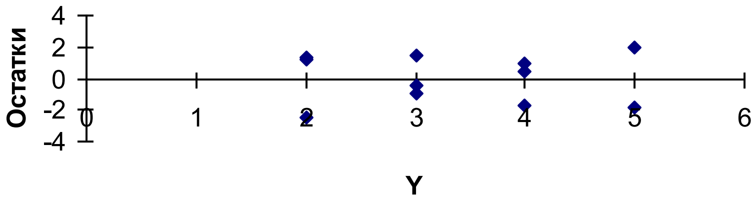


Рис. 5. График остатков регрессии

МУЛЬТИКОЛЛИНЕАРНОСТЬ

Увеличение числа переменных в уравнении множественной регрессии повышает точность описания взаимосвязи, однако при этом должно выполняться условие, что Xi – **объясняющие** переменные, линейно независимые величины.

Под **мультиколлинеарностью** понимают взаимосвязь объясняющих переменных регрессии. Если между переменными Xn и Xm существует функциональная зависимость (Xn= a Xm), то говорят о **строгой** мультиколлинеарности. Чаще всего между переменными существует довольно сильная корреляционная зависимость – в этом случае мультиколлинеарность называют **нестрогой**.

При строгой мультиколлинеарности решение матричного уравнения 1.22 становится невозможным, так как матрица X^T X вырожденная – её определитель равен нулю.

Если же мультиколлинеарность нестрогая, то решение матричного уравнения формально можно найти, однако все оценки мало надежны.

Чтобы обнаружить мультиколлинеарность надо найти определитель матрицы $X^T X$. Вместо этого проверяется определитель матрицы межфакторной корреляции, которую получают с помощью инструмента КОРРЕЛ.

Устранение мультиколлинеарности заключается в исключении одной из двух, находящихся во взаимосвязи переменных, либо путем пересмотра структуры уравнения регрессии. Для оценки влияния факторов на результирующий фактор Y в случае используются показатели частной корреляции (1.26). Если число переменных больше трех, то для их определения удобно пользоваться формулой:

$$r_{kp} = \frac{c_{kp}}{\sqrt{c_{kk} \cdot c_{pp}}},$$

где c_{kp} коэффициенты матрицы обратной к матрице парных коэффициентов корреляции.

ГОМОСКЕДАСТИЧНОСТЬ (постоянство дисперсии случайных отклонений)

Для применения МНК требуется, чтобы дисперсия остатков была величиной постоянной. Невыполнимость этого условия называется **гетероскедастичностью** и влечёт смещенность дисперсий оценок, так как стандартная ошибка регрессии (1.4) становится смещенной.

Обнаружение гетероскедастичности является сложной задачей потому что необходимо знать распределение СВY, соответствующее выбранному значению переменной x_i. В тесте Голфелда–Квандта предполагается, что стандартное отклонение пропорционально значению x_i переменной X и e_i нормально распределены, автокорреляция остатков отсутствует. Проверка на гомоскедастичность по этому тесту содержит следующие шаги:

1. Все n наблюдений упорядочивают по величине.
2. Упорядоченная выборка разбивается на три подвыборки размерностью k, (n-2k) и k соответственно.
3. Центральные наблюдения исключаются из дальнейшего рассмотрения.
4. Строят регрессии для первой и последней групп и находят остаточные суммы квадратов $S1^2$ и $S2^2$ соответственно.
5. Находят их отношение $F = \frac{S2^2}{S1^2}$. Если условие гомоскедастичности выполняется, то $S1^2 = S2^2$, в противном случае $S1^2 << S2^2$.
6. Построенная F-статистика, имеет распределение Фишера с $\nu_1 = \nu_2 = k-m-1$ степенями свободы, где m число объясняющих переменных в уравнении регрессии.
7. Чем больше F превышает значение $F_{кр}$, тем более нарушена предпосылка о равенстве остаточных дисперсий.

Раздел №6. Системы эконометрических уравнений.

При изучении сложных явлений часто сталкиваются с необходимостью статистического описания не отдельных зависимостей, а их комплекса. В этих случаях применяют системы уравнений регрессии, каждое из которых описывает одну статистическую закономерность. Как правило, развернутая математическая модель содержит кроме регрессионных уравнений уравнения, характеризующие балансовые связи между переменными или характеристики тенденций развития отдельных переменных. Так, если изучается модель спроса как соотношение цен и количества потребляемых товаров, то одновременно для прогнозирования спроса необходима модель предложения товаров, в которой рассматривается взаимосвязь между количеством и ценой предлагаемых благ, что позволяет достичь равновесия между спросом и предложением. Наибольшее использование системы взаимосвязных уравнений получили при макроэконометричекских расчетах.

Системы уравнений в эконометрических исследованиях могут быть нескольких видов.

Системы независимых уравнений. В этом случае каждая зависимая переменная y_i рассматривается как функция, зависящая от переменных из набора факторов $X=(x_1, x_2, ..x_n)$. Количество факторов x_i в каждом уравнении системы может быть различным. Отсутствие некоторого фактора в уравнении системы может быть следствием несущестщественности его воздействия на y_i так и экономической нецелесообразностью его включения в модель. Каждое уравнение системы может рассматриваться самостоятельно. Для нахождения параметров этого уравнения используется МНК, а фактическое значение зависимой переменной отличается от теоретического на величину случайной ошибки ε_i.

Системы рекурсивных уравнений. В таких системах зависимая переменная y_i включает в качестве факторов y_{i-1}, y_{i-2}, …., y_1 все предшествующие зависимые переменные и набор собственных независимых факторов x. Каждое уравнение может рассматриваться самостоятельно, и его параметры определяются МНК.

Системы взаимосвязанных уравнений. В них одни и те же зависимые переменные в одних уравнениях входят в левую часть, а в других – в правую часть системы. В отличие от предыдущих систем каждое уравнение не может рассматриваться самостоятельно, и для нахождения его параметров МНК нельзя применить. В эконометрике такая система уравнений называется также **структурной формой модели,** а для нахождения её параметров используются специальные методы.

Система совместных, одновременных уравнений или структурная форма модели содержит эндогенные и экзогенные переменные.

Эндогенные переменные это зависимые переменные, число которых равно числу уравнений системы. В системе уравнений они обозначаются как y.

Экзогенные переменные это независимые переменные их, как правило, обозначают как x, влияющие на эндогенные переменные.

Предопределенные переменные – экзогенные и лаговые (за предыдущие моменты времени) эндогенные переменны системы.

Простейшая структурная форма модели имеет вид:

$$\begin{aligned} y_1 &= \alpha_0 y_2 + \alpha_1 x_1 + \varepsilon_1 \\ y_2 &= \beta_0 y_1 + \beta_1 x_2 + \varepsilon_2. \end{aligned} \qquad 5.\ 1$$

Структурная форма модели позволяет увидеть влияние изменений любой экзогенной переменной на значение эндогенных переменных. По предположению система уравнений структурной формы модели должна быть совместной. Тогда она может быть решена относительно эндогенных переменных. Результат решения представляет собой **приведенную форму** модели. Для структурной формы (4.1)приведенная форма модели имеет вид:

$$\begin{aligned} y_1 &= \delta_{11} x_1 + \delta_{12} x_2 + \eta_1 \\ y_2 &= \delta_{21} x_1 + \delta_{22} x_2 + \eta_2\,, \end{aligned} \qquad 5.\ 2$$

где коэффициенты δ_{ij} получим из структурной формы модели (5.1) выполнив простейшие преобразования. Коэффициенты приведенной формы модели δ_{ij} являются нелинейными функциями коэффициентов структурной формы, а ошибки η_i являются линейными функциями возмущений ε_i.

В общем виде переход от структурной к приведенной форме модели можно представить используя матричную запись. Запишем структурную модель как

$$\Gamma y + Ax = \varepsilon \qquad 5.\ 3$$

где Г и А – матрицы неизвестных параметров размерности $n\times n$ и $n\times m$ соответственно; матрица Г – невырожденная; y,x – векторы экзогенных и эндогенных переменных; ε – вектор случайных нормально распределенных величин с нулевым математическим ожиданием; n – число эндогенных переменных; m – число экзогенных переменных.

Решив матричное уравнение (5.3) относительно y, получим

$$y = -\Gamma^{-1}Ax + \eta = Bx + \eta, \qquad 5.\ 4$$

$$\text{где} \quad B = -\Gamma^{-1}A. \qquad 5.\ 5$$

Уравнения приведенной формы с численными параметрами позволяют получить значения эндогенных переменных, однако они объясняют их только через экзогенные переменные, то есть с меньшей детализацией, чем структурная форма, так как отсутствуют взаимосвязи зависимых переменных между собой.

Необходимое условие идентификации систем взаимосвязанных уравнений.

Пусть n – полное число уравнений, что совпадает с числом эндогенных переменных, m – число экзогенных переменных всей системы, n_i+m_i – число эндогенных и экзогенных переменных в i-том уравнении. Тогда для идентифицируемости этого уравнения должно выполняться условие

$$\mathbf{n+m-(\ n_i+m_i)\geq n-1.} \qquad 5.6$$

Условие (5.6) является **необходимым** условием идентификации системы одновременных уравнений. То есть, число оставленных в полном наборе переменных не должно быть меньше числа уравнений системы без единицы.

1. *Модель считается идентифицируемой, если каждое уравнение системы идентифицируемо.*
2. *Если хотя бы одно уравнение неидентифицируемо, то и вся модель неидентифицируема..*
3. *Сверхидентифируемая модель содержит хотя бы одно сверхидентифицируемое уравнение.*

Выполнение условия идентифицуроемости модели проверяется для каждого уравнения системы. Для i-того уравнения надо проверить выполнение условия (5.6) и сделать вывод:

a) уравнение идентифицируемо, если между левой и правой частями в (5.6) достигается равенство;
b) уравнение сверхидентифицируемо, если левая часть неравенства строго больше правой в (5.6);
c) уравнение неидентифируемо, если неравенство (5.6) не выполняется.

Рассмотренное правило не является достаточным условием идентификации модели.

Достаточное условие идентифицикации систем уравнений выполняется, если определитель матрицы, составленной из коэффициентов исключенных из данного уравнения переменных (эндогенных и экзогенных), но входящих в другие уравнения системы не равен нулю и ранг этой матрицы не меньше **n-1**.

Пример 5. *Задание 1.* Идентификацируйте модель:

$$\begin{cases} y_1 = b_{13}y_3 + a_{11}x_1 + a_{13}x_3 \\ y_2 = b_{21}y_1 + b_{23}y_3 + a_{22}x_2 \\ y_3 = b_{32}y_2 + a_{31}x_1 + a_{33}x_3 \end{cases}$$

Модель имеет три эндогенные (y_1, y_2, y_3) и три экзогенные(x_1, x_2, x_3) переменные. Проверим для каждого уравнения необходимое и достаточное условия идентификации.

Первое уравнение.

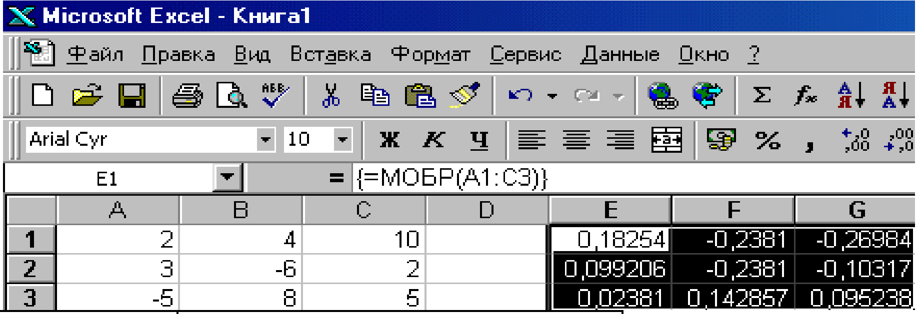
Необходимое условие: n_1=2, m_1=2, $n_1 + m_1$=4, равенство 6 – 4 =3-1 выполняется, следовательно уравнение точно идентифицируемо. Что бы проверить достаточное условие построим матрицу из коэффициентов при переменных y_2 и x_2 во втором и третьем уравнениях системы:

Уравнение	Отсутствующие переменные	
	y_2	x_2
второе	-1	a_{22}
третье	b_{32}	0

Определитель матрицы $\Delta A=-1\cdot 0-a_{22}\cdot b_{32}$; ранг матрицы равен 2 и достаточное условие идентификации выполняется. Уравнение точно идентифицировано.

Второе уравнение.

Необходимое условие выполняется: 6–4=3–1. Для проверки достаточного составим матрицу из коэффициентов отсутствующих переменных: x_1 и x_3.



Уравнение	Отсутствующие переменные	
	x_1	x_3
первое	a_{11}	a_{13}
третье	a_{31}	a_{33}

Определитель матрицы $\Delta A=a_{11}\cdot a_{33}-a_{13}\cdot a_{31}\neq 0$, ранг равен 2 – уравнение точно идентифицируемо.

Третье уравнение системы проверьте самостоятельно и убедитесь, что оно точно идентифицируемо.

Следовательно, вся модель точно идентифицируема.

Задание 2. Исходя из приведенной модели уравнений

$$\begin{cases} y_1 = 2x_1 + 4x_2 + 10x_3 \\ y_2 = 3x_1 - 6x_2 + 2x_3 \\ y_3 = -5x_1 + 8x_2 + 5x_3 \end{cases}$$

найдите структурные коэффициенты исходной модели. Используя встроенную функцию ППП Excel, найдем обратную матрицу. Решив матричное уравнение $A^{-1}Y=X$, мы можем записать уравнения выражающие x_i через y_i.

$$\begin{cases} x_1 = 0{,}1825y_1 - 0{,}2381y_2 - 0{,}2698y_3 \\ x_2 = 0{,}0993y_1 - 0{,}2381y_2 - 0{,}1032y_3 \\ x_3 = 0{,}02381y_1 + 0{,}14298y_2 + 0{,}095y_3 \end{cases}$$

Из первой системы, сложив все её уравнения получим: $y_1+y_2+y_3=30x_2+17x_3$. Подставив x_3 из второй системы в это уравнение и преобразовав его, получим второе уравнение структурной формы модели $y_1+y_2+y_3=6x_2+17\cdot(0{,}024y_1+0{,}143y_2+0{,}095y_3)$. Выполним приведение подобных и уравнение примет вид:

$$y_2= 0{,}416y_1-0{,}434\, y_3-4{,}2x_2.$$

Исключим x_2, сложив третье и второе уравнения приведенной формы:

$$\begin{array}{l} y_2 = 3x_1 - 6x_2 + 2x_3 \,|\cdot 8 \\ \underline{y_3 = -5x_1 + 8x_2 + 5x_3 \,|\cdot 6} \\ 8y_2 + 6y_3 = -6x_1 + 46x_3 \end{array}.$$

Выразим из этого уравнения y_3 и получим третье уравнение структурной формы: $y_3=-1{,}336\, y_2-x_1-7{,}664\, x_3$.

Чтобы получить первое уравнение структурной формы умножим первое уравнение на –2 и сложим его с третьим уравнением приведенной формы (выполните самостоятельно). После преобразования получим: $y_1=0{,}5\, y_3+4{,}5x_1+7{,}5\, x_3$.

Таким образом, структурная форма модели примет вид:

$$\begin{cases} y_1 = 0{,}5y_3 + 4{,}5x_1 + 7{,}5x_3 \\ y_2 = 0{,}416y_1 - 0{,}434y_3 - 4{,}2x_2 \\ y_3 = -1{,}336y_2 - x_1 + 7{,}664x_3 \end{cases}.$$

ОЦЕНИВАНИЕ ПАРАМЕТРОВ СТРУКТУРНОЙ МОДЕЛИ

МНК для рекурсивных моделей.

В таких моделях эндогенные переменные последовательно (рекурсивно) связаны друг с другом. Первая эндогенная переменная y_1 зависит только от экзогенных переменных x_i, $i=1, 2, \ldots m$ и случайного отклонения ε_1. Вторая эндогенная переменная y_2 зависит от экзогенных переменных x_i, $i=1, 2, \ldots m$ и случайного отклонения ε_2, а также от предшествующей эндогенной переменной y_1 и т. д.

В этих моделях структурные уравнения оцениваются поэтапно начиная с первого ($y_1 \to y_2 \to \ldots \to y_n$). Оценки полученные с применением МНК являются несмещенными и состоятельными.

Однако модели такого типа встречаются редко. В общем случае для оценки коэффициентов структурного уравнения сначала необходимо преобразовать исходное уравнение к приведенному виду, а затем применить обыкновенный метод наименьших квадратов. Основанный на данной процедуре метод называется косвенный метод наименьших квадратов (КМНК).

КОСВЕННЫЙ МЕТОД НАИМЕНЬШИХ КВАДРАТОВ (КМНК).

Процедура применения КМНК предполагает выполнения следующих этапов:

a) структурная модель преобразуется в приведенную форму;
b) для каждого уравнения приведенной формы модели обычным МНК оцениваются приведенные коэффициенты (δij);
c) по коэффициентам приведенной формы модели находим параметры исходного уравнения.

ДВУХШАГОВЫЙ МЕТОД НАИМЕНЬШИХ КВАДРАТОВ (ДМНК).

Если система сверхидентифицируема, то КМНК не используется, так как он не дает однозначных оценок параметров. При его применении надо выполнить следующие этапы:

a) структурная модель преобразуется в приведенную форму;
b) для каждого уравнения приведенной формы модели обычным МНК оцениваются приведенные коэффициенты (δij);
c) получить теоретические значения $\hat{y}i$, содержащихся в правой части уравнения;
d) подставить новые значения вместо фактических и применить МНК к структурной форме модели.

Если все уравнения системы сверхидентифицируемые, то ДМНК применяется каждому уравнению. Если в системе есть точно идентифицируемые уравнения, то структурные коэффициенты для них находятся из системы приведенных уравнений. ДМНК является наиболее распространенным методом решения систем одновременных уравнений, поэтому он используется в ряде компьютерных программ и для точно идентифицируемых уравнений, например DSTAT, однако при использовании этой программ необходимо помнить, что она не работает, если в модели вводятся ограничения на параметры.

Пример 6. Рассмотрим на примере модели вида (5.1)

$$y_1=\alpha_0 y_2 + \alpha_1 x_1 + \varepsilon_1$$
$$y_2=\beta_0 y_1 + \beta_1 x_2 + \varepsilon_2.$$

ПРИМЕНЕНИЕ КМНК.

Шаг 1.Пусть для построения данной модели мы располагаем некоторой информацией по пяти регионам:

Регион	y_1	y_2	x_1	x_2
1	2	5	1	3

2	3	6	2	1
3	4	7	3	2
4	5	8	2	5
5	6	5	4	6
Средние	4	6,2	2,4	3,4

Приведенная форма модели имеет вид (4.2):

$$y_1=\delta_{11}x_1 + \delta_{12}x_2 + \eta_1$$
$$y_2=\delta_{21}x_1 + \delta_{22}x_2 + \eta_2.$$

Для каждого уравнения приведенной формы определим δ-коэффициенты, используя Анализ данных ППП Excel.

Для первого уравнения получим. Используя данные рис.10 можем записать первое уравнение приведенной формы:

$$y_1= 0,852\ x_1+0,373\ x_2 +\eta_1.$$

Для второго уравнения поступим аналогичным образом и получим:

$$y_2= –0,072\ x_1–0,006\ x_2+\eta_2.$$

Таким образом, приведенная форма модели:

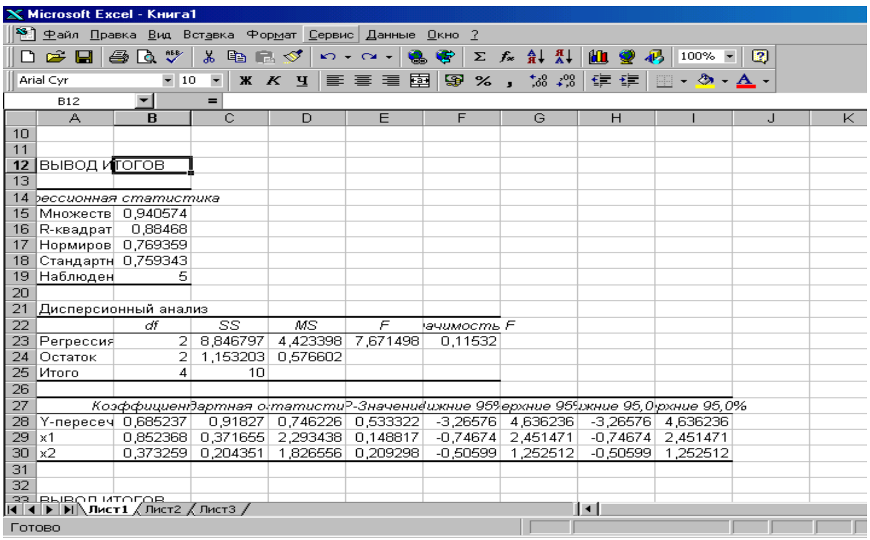
$$y_1= 0,852\ x_1+0,373\ x_2 +\eta_1$$
$$y_2= –0,072\ x_1–0,006\ x_2+\eta_2.$$

Шаг 2: перейдем к структурной модели. Для этого можно используя встроенную функцию МОБР(массив), выразим x_1 и x_2 через y_1 и y_2.

$$x_1= –0,25\ y_1 –16,75\ y_2 +\varepsilon_1 \qquad 5.7$$
$$x_2= 3,25y_1 +38,25\ y_2+\varepsilon_2. \qquad 5.8$$

Подставив в первое уравнение приведенной формы значение x_2 из (5.8), а во второе x_1 из (5.9) получим структурную форму модели:

$$y_1=67,029y_2 + 4x_1 + \varepsilon_1$$
$$y_2=–0,085y_1 + 0,028x_2 + \varepsilon_2.$$



	A	B	C	D	E	F	G	H	I
10									
11									
12	ВЫВОД ИТОГОВ								
13									
14	рессионная статистика								
15	Множеств	0,940574							
16	R-квадрат	0,88468							
17	Нормиров	0,769359							
18	Стандартн	0,759343							
19	Наблюден	5							
20									
21	Дисперсионный анализ								
22		df	SS	MS	F	ачимость F			
23	Регрессия	2	8,846797	4,423398	7,671498	0,11532			
24	Остаток	2	1,153203	0,576602					
25	Итого	4	10						
26									
27		Коэффициен	дартная о	татисти	Р-Значени	ижние 95%	ерхние 95%	ижние 95,0	рхние 95,0%
28	Y-пересеч	0,685237	0,91827	0,746226	0,533322	-3,26576	4,636236	-3,26576	4,636236
29	x1	0,852368	0,371655	2,293438	0,148817	-0,74674	2,451471	-0,74674	2,451471
30	x2	0,373259	0,204351	1,826556	0,209298	-0,50599	1,252512	-0,50599	1,252512
31									
32									
33	ВЫВОД ИТОГОВ								

Рис. 10. Вывод итоговой регрессии для первого уравнения приведенной формы.

Эту же систему можно записать, включив в нее свободный член уравнений, для этого надо перейти от переменных в виде отклонений от среднего уровня к исходным переменным y и x.

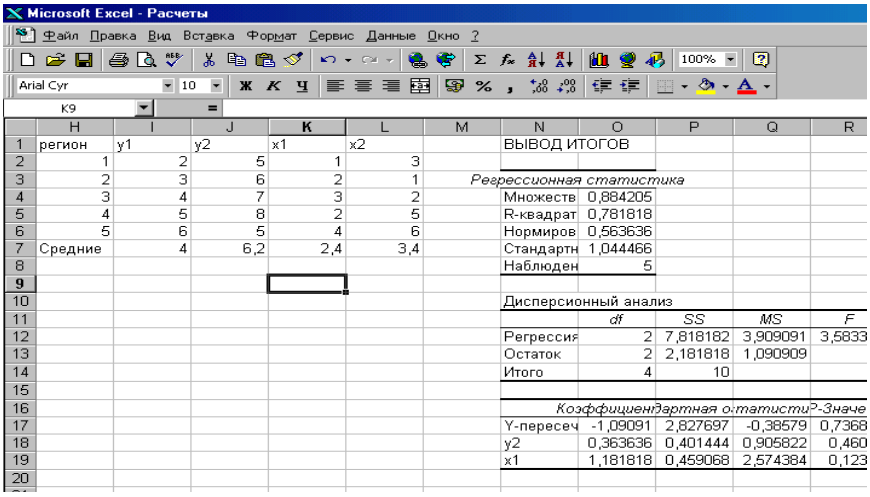
Свободные члены уравнений определим по формулам:

$$a_1= \bar{y}_1 –b_{12}\ \bar{y}_2 –a_{11}\ \bar{x}_1=429; \quad a_2= \bar{y}_2 –b_{21}\ \bar{y}_1 –a_{22}\ \bar{x}_2 =6,45.$$

Тогда структурная модель примет вид:

$$y_1=429 + 67,029y_2 + 4x_1 + \varepsilon_1$$

$$y_2=6{,}45–0{,}085y_1 + 0{,}028x_2 + \varepsilon_2.$$



	H	I	J	K	L	M	N	O	P	Q	R
1	регион	y1	y2	x1	x2		ВЫВОД ИТОГОВ				
2	1	2	5	1	3						
3	2	3	6	2	1		*Регрессионная статистика*				
4	3	4	7	3	2		Множеств	0,884205			
5	4	5	8	2	5		R-квадрат	0,781818			
6	5	6	5	4	6		Нормиров	0,563636			
7	Средние	4	6,2	2,4	3,4		Стандартн	1,044466			
8							Наблюден	5			
9											
10							Дисперсионный анализ				
11								*df*	*SS*	*MS*	*F*
12							Регрессия	2	7,818182	3,909091	3,5833
13							Остаток	2	2,181818	1,090909	
14							Итого	4	10		
15											
16								*Коэффициент*	*дартная о*	*татисти*	*Р-Значе*
17							Y-пересеч	-1,09091	2,827697	-0,38579	0,7368
18							y2	0,363636	0,401444	0,905822	0,460
19							x1	1,181818	0,459068	2,574384	0,123
20											

Рис. 11. Коэффициенты первого уравнения множественной регрессии, найденные с помощью диалогового окна регрессия.

Оценка качества модели делается с помощью F– теста и R^2 для каждого уравнения в отдельности. Для первого уравнения $F_{набл.}=7{,}67$ (найдите его на рис. 8), а $F_{табл.}=19$ при уровне значимости $\alpha=0{,}05$, что говорит о невысоком качестве этого уравнения. Однако коэффициент детерминации $R^2=0{,}94$ очень высокий и можно предположить, что выбрано мало наблюдений.

Если к каждому уравнению структурной формы применить МНК, то результат может сильно отличаться от полученного по КМНК, например, для первого уравнения системы получится:

Коэффициенты при переменных, полученных по МНК (рис. 11), сильно отличаются от соответствующих коэффициентов полученных в структурной форме для первого уравнения.

Надо помнить, что МНК, применяемый к каждому уравнению структурной формы модели, взятому в отдельности, дает смещенные оценки структурных коэффициентов. Доказано, что коэффициент регрессии отличается от структурного коэффициента и совпадает с ним только в случае, когда переменная y не содержит ошибок (т.е. $\varepsilon_i=0$), а ошибки x имеют дисперсию равную нулю.

Пример 6. Рассмотрим применение ДМНК на модели вида:

$$y_1=\alpha_1(y_2 + x_1) + \varepsilon_1$$
$$y_2=\beta_0y_1 + \beta_1x_2 + \varepsilon_2.$$

Эта модель может быть получена из предыдущей идентифицируемой модели:

$$y_1=\alpha_0y_2 + \alpha_1x_1 + \varepsilon_1$$
$$y_2=\beta_0y_1 + \beta_1x_2 + \varepsilon_2,$$

если наложить ограничения на её параметры $\alpha_1=\alpha_0$. В результате первое уравнение стало сверхидентифицируемым: 4-2> 2-1. Второе уравнение осталось точно идентифицируемым.

Шаг 1: получим приведенную форму модели и найдем её коэффициенты δ_{ij}.

Для этого воспользуемся результатами предыдущего примера

$$y_1= 0{,}852\ x_1+0{,}373\ x_2 +\eta_1$$
$$y_2= –0{,}072\ x_1–0{,}006\ x_2+\eta_2. \qquad 5.9$$

Шаг 2: найдем теоретические значения $\hat{y}_2$ для эндогенной переменной y_2 . Для этого в уравнение (5.9) подставим значения $x_1-\bar{x}$ и $x_2-\bar{x}$ из таблицы 2 и вычислим значение y_2

– $\bar{y}$ (расчеты выполним в Excel). Расчетные данные для второго шага ДМНК приведены в таблице 2.

Шаг 3: теперь используя встроенную функцию ЛИНЕЙН (Знач У; Знач Х; Ложь;) по данным таблицы 2. найдем параметр α_1=1,243 структурного уравнения $y_1= \alpha_1(y_2 + x_1)$.

Таким образом, первое структурное уравнение примет вид:

$$y_1= 1{,}243\cdot(y_2 + x_1).$$

Шаг 4: так как второе уравнение системы точно идентифицируемое, то воспользуемся результатами примера 5: $y_2=–0{,}085y_1 + 0{,}028x_2 + \varepsilon_2$.

В целом система одновременных уравнений примет вид:

$$y_1= 1{,}243\cdot(y_2 + x_1)$$
$$y_2=–0{,}085y_1 + 0{,}028x_2 + \varepsilon_2.$$

Таблица 2

Microsoft Excel - Расчеты

	A	B	C	D	E	F
11	Регион	x1-x1ср	x2-x2cp	y2т = -0,072 x1 -0,005 x2	y2т+x1	y1-yср
12	1	-1,4	-0,4	0,1032	-1,2968	-2
13	2	-0,4	-2,4	0,0432	-0,3568	-1
14	3	0,6	-1,4	-0,0348	0,5652	0
15	4	-0,4	1,6	0,0192	-0,3808	1
16	5	1,6	2,6	-0,1308	1,4692	2
17	Сумма	0	0	0	0	0

Раздел №7. Одномерные временные ряды

Почти в каждой области встречаются явления, которые интересно и важно изучать в их развитии и изменении во времени. В повседневной жизни могут представлять интерес, например, метеорологические условия, цены на тот или иной товар, те или иные характеристики состояния здоровья индивидуума и т. д. Все они изменяются во времени. С течением времени изменяются деловая активность, режим протекания того или иного производственного процесса, глубина сна человека, восприятие телевизионной программы. Совокупность измерений какой-либо одной характеристики подобного рода в течение некоторого периода времени представляют собой временной ряд.

Совокупность существующих методов анализа таких рядов наблюдений называется анализом временных рядов.

Основной чертой, выделяющей анализ временных рядов среди других видов статистического анализа, является существенность порядка, в котором производятся наблюдения. Если во многих задачах наблюдения статистически независимы, то во временных рядах они, как правило, зависимы, и характер этой зависимости может определяться положением наблюдений в последовательности. Природа ряда и структура порождающего ряд процесса могут предопределять порядок образования последовательности.

Цель работы состоит в получении модели для дискретного временного ряда во временной области, обладающей максимальной простотой и минимальным числом параметров и при этом адекватно описывающей наблюдения.

Получение такой модели важно по следующим причинам:

1) она может помочь понять природу системы, генерирующей временные ряды;
2) управлять процессом, порождающим ряд;
3) ее можно использовать для оптимального прогнозирования будущих значений временных рядов;

Временные ряды лучше всего описываются нестационарными моделями, в которых тренды и другие псевдоустойчивые характеристики , возможно меняющиеся во времени , рассматриваются скорее как статистические, а не детерминированные явления. Кроме того, временные ряды, связанные с экономикой , часто обладают заметными *сезонными*, или периодическими , компонентами; эти компоненты могут меняться во времени и должны описываться циклическими статистическими (возможно, нестационарными) моделями.

Пусть наблюдаемым временным рядом является $y_1, y_2, \ldots, y_n$. Мы будем понимать эту запись следующим образом. Имеется T чисел, представляющих собой наблюдение некоторой переменной в T равноотстоящих моментов времени. Эти моменты для удобства пронумерованы целыми числами 1, 2, . . .,T. Достаточно общей математической (статистической или вероятностной) моделью служит модель вида:

$$y_t = f(t) + u_t\ ,\ t = 1, 2, \ldots, T.$$

В этой модели наблюдаемый ряд рассматривается как сумма некоторой полностью детерминированной последовательности $\{f(t)\}$, которую можно назвать математической составляющей, и случайной последовательности $\{u_t\}$, подчиняющейся некоторому вероятностному закону. (И иногда для этих двух составляющих используются соответственно термины сигнал и шум). Эти компоненты наблюдаемого ряда ненаблюдаемы; они являются теоретическими величинами. Точный смысл указанного разложения зависит не только от самих данных, но частично и оттого, что понимается под повторением эксперимента, результатом которого являются эти данные. Здесь используется так называемая «частотная» интерпретация. Полагается, что, по крайней мере, принципиально можно повторять всю ситуацию целиком, получая новые совокупности наблюдений. Случайные составляющие , кроме всего прочего, могут включать в себя ошибки наблюдений.

В данной работе рассмотрена модель временного ряда, в которой на тренд накладывается случайная составляющая, образующая случайный стационарный процесс. В такой модели предполагается, что течение времени никак не отражается на случайной составляющей. Точнее говоря, предполагается, что математическое ожидание (то есть среднее значение) случайной составляющей тождественно равно нулю, дисперсия равна некоторой постоянной и что значения u_t в различные моменты времени некоррелированны. Таким образом, всякая зависимость от времени включается в систематическую составляющую f(t). Последовательность f(t) может зависеть от некоторых неизвестных коэффициентов и от известных величин, меняющихся со временем. В этом случае её называют «функцией регрессии». Методы статистических выводов для коэффициентов функции регрессии оказываются полезными во многих областях статистики. Своеобразие же методов, относящихся именно к временным рядам, состоит в том, что здесь исследуются те модели, в которых упомянутые выше величины, меняющиеся со временем, являются известными функциями t.

1. Основные элементы временного ряда

1.1. Основные понятия

Изучение изменения объекта начинается с построения таблиц, в которых представлены количественные показатели признака этого объекта за последовательные моменты времени. Эти данные называют рядами динамики или временными рядами. Каждое конкретное значение называется уровнем ряда. Каждое значение временного ряда формируется под воздействием различных факторов, которые условно можно разделить на три группы:

1. факторы, формирующие тенденцию ряда (T);
2. факторы, формирующие циклические колебания ряда (S);
3. случайные факторы (E).

Модель, в которой временной ряд представляется как сумма, перечисленных компонент называется аддитивной (Y = T + S +E).

Модель, в которой временной ряд представляется как произведение, перечисленных компонент называется мультипликативной (Y = T · S · E)..

Основной задачей эконометрического исследования временных рядов является выявление тенденции, цикличности и случайной компоненты временного ряда и описание их количественного выражении.

1.2. Показатели, характеризующие тенденцию временного ряда

Определяют величину изменения уровня ряда, как в абсолютном, так и в относительном выражении (на какую долю, процент от базового уровня произошло изменение). Интересно являются ли эти изменения равномерными или непрерывными, ускоренными или замедленными и самое главное каким уравнением их можно описать? Впервые понятие об уравнении тенденции ряда было введено английским ученым Гукером в 1902 году. Он предложил называть такое уравнение трендом. При составлении этого уравнения следует представить временной ряд в форме «чистого» тренда при отсутствии колебаний.

Абсолютное изменение уровней характеризуется разностью между сравниваемым уровнем и уровнем, принятым за базисный. В качестве базы может быть выбран предыдущий уровень и тогда полученный показатель называется цепным, или начальный (нулевой уровень), тогда говорят о базовом показателе.

Формулы для абсолютного изменения уровня:

1. цепной: $\Delta_i = y_i - y_{i-1}$; (1.1)
2. базисный: $\Delta_{0i} = y_i - y_0$. (1.2)

Ускорением называют разность между абсолютными изменениями за данный период и абсолютным изменением за предшествующий период одинаковой длительности:

$$\Delta_i^{/} = \Delta_i - \Delta_{i-1}. \qquad (1.3)$$

Показатель ускорения применяется только в цепном варианте, но не в базисном. Если он принимает отрицательное значение, то говорят о замедлении роста. В случае положительных значений этого показателя можно говорить о возрастании роста показателя.

Относительные показатели характеризуют темпы роста временного ряда.

Обозначим темп изменения через k. Тогда:

Цепной темп роста : $k_{i/i-1} = \dfrac{y_i}{y_{i-1}}$; (1.4)

Базисный темп роста: $k_{i/0} = \dfrac{y_i}{y_0}$. (1.5)

Как правило, показатели темпа роста измеряются в процентах. Рассмотрим связь между абсолютными и относительными и показателями динамики. Выразим сравниваемый уровень через уровень предыдущего года плюс абсолютный прирост, тогда абсолютный показатель:

$$k_{i/i-1} = \frac{y_{i-1} + \Delta_i}{y_{i-1}} = 1 + \frac{\Delta_i}{y_{i-1}} \text{ или } k_{i/i-1}(\%) = 100\%(1 + \frac{\Delta_i}{y_{i-1}}) \qquad (1.6)$$

$$k_{i/0} = \frac{y_0 + \Delta_{0i}}{y_0} = 1 + \frac{\Delta_{0i}}{y_0} \text{ или } k_{i/0}(\%) = 100\%(1 + \frac{\Delta_{0i}}{y_0}). \qquad (1.7)$$

Величину $\dfrac{\Delta_i}{y_{i-1}}$ или $\dfrac{\Delta_{0i}}{y_0}$ называют относительным приростом или темпом прироста. Если значения уровней ряда принимают как положительные, так и отрицательные значения, то темп роста и темп прироста для исследования временного ряда применять нельзя.

Рассмотрим соотношения между цепными и базисными показателями на примере данных из таблицы 1.

1.Сумма цепных абсолютных изменений равна базисному абсолютному изменению:

$$\sum_{n=0}^{i} \Delta_n = \Delta_{0i}.$$

2.Произведение цепных темпов роста равно базисному темпу роста:

$$\prod_{n=1}^{i} k_{n/n-1} = k\,i/0.$$

3.Складывать темпы прироста нельзя.

Пример 1.1. По данным таблицы 1 найдите показатели, характеризующие динамику временного ряда.

Таблица 1

Относительные и абсолютные показатели тенденции продаж риса в Китае

Номер периода	Уровень ряда	Абсолютное изменение	Ускорение	Темп роста в % к предыдущему	Темп роста к начальному, в %
0	591.65				
1	594.58	2.93		100.4952252	100.4952
2	594.81	0.23	-2.7	100.0386828	100.5341
3	592.81	-2	-2.23	99.66375817	100.1961
4	598.25	5.44	7.44	100.9176633	101.1155
5	602.34	4.09	-1.35	100.6836607	101.8068
6	601.41	-0.93	-5.02	99.84560215	101.6496
7	597.06	-4.35	-3.42	99.27669976	100.9144
8	599.07	2.01	6.36	100.3366496	101.2541
9	597.6	-1.47	-3.48	99.75461966	101.0057

ШАГ 1. Откройте книгу Excel и введите заголовки столбцов и значения первых двух столбцов. Отформатируйте их как в таблице 1.

ЩАГ 2. Вставьте формулы:

в ячейку C3 : = B3-B2;

в ячейку D4: = =C4-C3;

в ячейку E3: = B3/B2%;

в ячейку F3: = B3/$B$2 %;

скопируйте их в остальные ячейки таблицы.

ШАГ 3. Проверьте соотношения между цепными и базисными показателями.

1.3. Средние показатели тенденции динамики

Средние значения необходимы для обобщения характеристик тенденции за длительный период. С их помощью можно сравнивать развитие показателя за неодинаковые по длительности отрезки времени, при выборе аналитического выражения для тренда. К средним показателям динамики относятся:

– средний уровень ряда $\bar{y} = \dfrac{\sum_{i=1}^{n} y_i}{n}$;

– средний абсолютный прирост $\bar{\Delta} = \dfrac{\sum \Delta_i}{n} = \dfrac{y_n - y_0}{n}$;

– среднее ускорение абсолютного изменения находится после аналитического выравнивания;

– средний темп роста $\bar{k} = \sqrt[n]{\prod_{i=1}^{n} k_i} = \sqrt[n]{\frac{y_n}{y_0}}$.

Для правильной интерпретации средних показателей, надо указать две единицы времени:

1) время, за которое он вычислен и которое он характеризует;
2) время, входящее в его единицу измерения.

Можно рассчитать среднемесячный прирост за пятилетие, среднесуточное изменение за год, месяц, квартал; среднемесячный темп за полугодие и т.д.

Пример 1.2. имеются данные о величине дохода (x) на одного члена семьи и расхода (y) на товар А (таблица 1.2).

Таблица 1.2

Год	Расходы на товар , доллар. (yi)	Доход на одного члена семьи, % (xi)	Δyi	Δxi
2000	147.9125	591.65		
2001	148.645	594.58	0.7325	2.93
2002	148.7025	594.81	0.0575	0.23
2003	148.2025	592.81	-0.5	-2
2004	149.5625	598.25	1.36	5.44
2005	150.585	602.34	1.0225	4.09
2006	150.3525	601.41	-0.2325	-0.93
2007	149.265	597.06	-1.0875	-4.35
2008	149.7675	599.07	0.5025	2.01
2009	149.4	597.6	-0.3675	-1.47

Требуется:

1. Определить ежегодные абсолютные приросты доходов и расходов. Сделать выводы о тенденции развития каждого ряда.
2. Определите ускорение роста доходов и расходов. Постройте графики, позволяющие провести сравнение со средним значением ускорения. Можно ли сделать выводы.
3. Определите темпы роста для x и y проведите графическое сравнение этих показателей с их средними значениями.

Рис. 1.1. График абсолютного прироста для изучения расходов семей на приобретение товара А

Абсолютные приросты находятся по формуле (1.1). Абсолютные приросты по у не имеют четко выраженной направленности.

2. Автокорреляция и авторегрессия

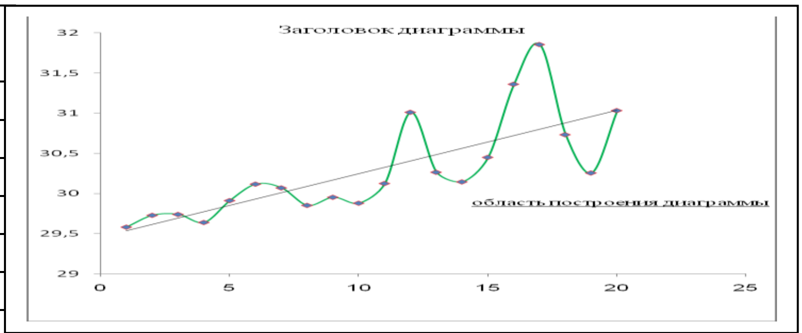
Рассмотрим временные последовательности не имеющие сезонных колебаний, при этом соседние наблюдения обычно близки, а достаточно далекие могут сильно отличаться. Такой плавающий рисунок характерен для многих последовательных данных в экономике, например, как цены акций в экономике.

Первым шагом является исследование графика временной последовательности. Исходя из графика, можно предположить вид регрессионной зависимости и исследовать ее с помощью инструментов регрессии.

2.1. Моделирование тенденции временного ряда с использованием Excel

Необходимо построить функциональную зависимость уровней временного ряда от времени. Наиболее простым способом является Добавление тренда к точечному графику, построенному по точкам (t, y_t). Воспользуйтесь данными о зарплате рабочих рис. (2. 1) для построения моделей.

Месяц	Зарплата , доллар
1	29.5825
2	29.729
3	29.7405
4	29.6405
5	29.9125
6	30.117
7	30.0705
8	29.853
9	29.9535
10	29.88
11	30.125
12	31.011
13	30.265
14	30.145
15	30.45
16	31.36
17	31.854
18	30.732



19	30.256
20	31.032

Рис 2.1. Данные о зарплате и график временного ряда

Построим линейный тренд по данным о зарплате и исследуем это приближение с помощью инструментов регрессии. Результаты (рис 2.2) показывают хорошее приближение к наблюдаемым данным. Значение R-квадрат показывает, что приблизительно 73% теоретических значений, вычисленных по уравнению регрессии $\hat{y} = 5{,}7768 + 0{,}0098x$, совпадает с наблюдаемыми. Величина t – статистик и P – значений подтверждают существенную значимость, полученного уравнения.

Регрессионная статистика	
Множественный R	0.7611045
R-квадрат	0.5792801
Нормированный R-квадрат	0.5559068
Стандартная ошибка	0.4092016
Наблюдения	20

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересечение	29.455982	0.190087	154.9605	1.377E-29
Переменная X 1	0.078997	0.0158682	4.978334	9.73E-05

Рис. 2.2. Результаты линейной регрессии

2.2.Проверка выполнимости МНК. Автокорреляция остатков Статистика Дарбина–Уотсона

Все предыдущие рассуждения основаны на том, что выполняются предпосылки МНК: предполагалось, что случайные отклонения e_i являются независимыми случайными величинами со средней, равной нулю. При работе с фактическими данными, такое допущение не всегда выполняется. Например, если вид функции выбран неудачно, то отклонения от регрессии вряд ли будут независимыми. В этом случае замечается концентрация положительных или отрицательных отклонений от регрессии и можно сомневаться в их случайном характере.

Если последовательные значения e_i коррелируют (зависят) между собой, то говорят, что имеет место автокорреляция остатков.

МНК в случае автокорреляции дает несмещенные и состоятельные оценки, однако полученные в этом случае доверительные интервалы имеют мало смысла в силу своей ненадежности. Значительная автокорреляция говорит о том, что спецификация модели неправильная. Проверка остатков на автокорреляцию должна выполняться обязательно. Наиболее простым приемом обнаружения автокорреляции является метод Дарбина–Уотсона (DW). Идея, которого состоит в том, что проверяются на коррелированность не любые, а только соседние значения e_i. Соседними обычно считаются соседние по возрастанию объясняющей переменной X (в случае перекрестной выборки) или по времени (в случае временных рядов) значения e_i.

Статистика DW рассчитывается по формуле:

$$d = DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \qquad 2.1$$

При условии что $M(e_i)=0$, $(i=1,2,\dots n)$ и n большое число можно предположить $\sum_{i=1}^{n} e_i^2 \approx \sum_{i=2}^{n} e_{i-1}^2$

Тогда после преобразования получим:

$$d = \frac{2\left(\sum_{i=2}^{n} e_i^2 - \sum_{i=2}^{n} e_i e_{i-1}\right)}{\sum_{i=1}^{n} e_i^2} = 2 - 2\frac{\sum_{i=2}^{n} e_i e_{i-1}}{\sum_{i=1}^{n} e_i^2} \qquad 2.2$$

Очевидно, что $0 \le d \le 4$, так как коэффициент корреляции $\left|r_{e_i e_{i-1}}\right| \le 1$:

d=2, если $cov(e_i, e_{i-1})=0$ – автокорреляция отсутствует;

d=0 –полная положительная автокорреляция;

d=4 –полная отрицательная автокорреляция.

2.3. Автокорреляция уровней временного ряда и выявление его структуры

Обычно проводится исследование корреляции временного ряда с самим собой до моделирования. При наличии тенденции и циклических колебаний значения каждого последующего уровня ряда зависят от предыдущих значений. Корреляционную зависимость между уровнями временного ряда или отклонениями от тренда, взятыми со сдвигом во времени называют автокорреляцией уровней ряда. В зависимости от величины сдвига во времени говорят о коэффициентах автокорреляции первого порядка (r1), второго порядка (r2) и т.д.

В зарубежной литературе предыдущие значения называют запаздывающими, а длина временного промежутка между текущим и предыдущим значениями называется временем запаздывания. В отечественной литературе время запаздывания называется лагом.

Пример. По данным о зарплате (рис. 2.1) в некоторой отрасли за 17 месяцев. Определите коэффициент автокорреляции первого порядка.

Месяц	Запаздывание1	Зарплата , доллар
1		29.5825
2	29.5825	29.729
3	29.729	29.7405
4	29.7405	29.6405
5	29.6405	29.9125
6	29.9125	30.117
7	30.117	30.0705
8	30.0705	29.853
9	29.853	29.9535
10	29.9535	29.88
11	29.88	30.125
12	30.125	31.011
13	31.011	30.265
14	30.265	30.145
15	30.145	30.45
16	30.45	31.36
17	31.36	31.854
18	31.854	30.732

19	30.732	30.256
20	30.256	31.032

Рис. 2.3. Подготовка данных для изучения автокорреляции с запаздыванием 1 (лагом 1)

Шаг 1. Введите данные о месяцах и зарплатах в столбцы A и B.

Шаг 2. выделите ячейку B1 и вызовите контекстное меню с помощью правой кнопки мыши. Выберите Добавить ячейки (Столбец).

Шаг 3. Скопируйте данные диапазона C2:C17 в ячейку B3. В ячейку B1 введите метку Запаздывание 1.

Щаг4. Удалите вторую строку.

Шаг 5. В ячейку E1 введите формулу:

=КОРРЕЛ(B2:B18; C2:C18).

Шаг 6. Постройте точечную диаграмму для данных из диапазона B2:C18.

Шаг 7. Сделайте вывод о зависимости между текущей зарплатой и предшествующей ей.

Отметим два важных свойства коэффициента автокорреляции:

Рассматривается линейная зависимость между текущим и предыдущим уровнями ряда. Поэтому по его значению можно сделать вывод только о линейной зависимости между рассматриваемыми данными. В примере r1=0,8400, что говорит о линейной зависимости между уровнями ряда с запаздыванием 1. Если существует нелинейная тенденция между уровнями ряда, то коэффициент корреляции может равняться нулю.

По знаку коэффициента автокорреляции нельзя делать вывод о возрастающей или убывающей тенденции.

Последовательность коэффициентов автокорреляции первого, второго и т.д. порядков называют функцией автокорреляции (ФА) временного ряда. График зависимости её значений от величины лага называется коррелограммой. Эти характеристики используют для выявления во временном ряде трендовой и циклической компонент.

Анализ автокорреляционной функции и коррелограммы позволяют определить лаг (запаздывание) при котором автокорреляция наибольшая, то есть связь между текущими и предшествующими значениями наиболее тесная.

Выводы о структуре временного ряда делают по правилу:

- Если самым большим оказался коэффициент автокорреляции первого порядка, исследуемый ряд содержит только тенденцию (T);
- Если самым большим оказался коэффициент к-того порядка, то ряд содержит циклические колебания (S) с лагом в к моментов времени;
- Если ни один из коэффициентов не является значимым, то возможно ряд содержит сильную нелинейную тенденцию или ряд зависит только от случайной компоненты (E).

Пример. По данным о зарплате вычислите функцию автокорреляции временного ряда и сделайте вывод о структуре исследуемых данных.

Для построения функции автокорреляции необходимо исследовать ряды с запаздываниями. На рис. 2.3 представлены данные для изучения коэффициента корреляции первого порядка. Можно аналогично подготовить таблицу данных (рис. 2.4). Для каждого запаздывания придется вычислять функцию

КОРРЕЛ(диапазон1;диапазон2), как показано на рис.2.4 столбец H.

Существует другой способ, позволяющий автоматизировать этот процесс. В столбце I (рис.2.4) для вычисления r1, r2,…, r6 используется функция:

СМЕЩ(диапазон; лаг_по строкам;лаг_по столбцам; ширина; высота).

Месяц	Лаг3	Лаг2	Лаг1	Зарплата , доллар	Лаг	ФА	ФА с функцией СМЕЩ

4	29.5825	29.729	29.7405	29.6405	1	0.591632	0.999388
5	29.729	29.7405	29.6405	29.9125	2	0.558492	0.998649
6	29.7405	29.6405	29.9125	30.117	3	0.524824	0.996687
7	29.6405	29.9125	30.117	30.0705	4	0.507291	0.994042
8	29.9125	30.117	30.0705	29.853	5	0.492262	0.969426
9	30.117	30.0705	29.853	29.9535	6	0.463714	0.965572
10	30.0705	29.853	29.9535	29.88			
11	29.853	29.9535	29.88	30.125			
12	29.9535	29.88	30.125	31.011			
13	29.88	30.125	31.011	30.265			
14	30.125	31.011	30.265	30.145			
15	31.011	30.265	30.145	30.45			
16	30.265	30.145	30.45	31.36			
17	30.145	30.45	31.36	31.854			
18	30.45	31.36	31.854	30.732			
19	31.36	31.854	30.732	30.256			
20	31.854	30.732	30.256	31.032			

Рис. 2.4. Лист Excel с данными для вычисления функции автокорреляции

В этом случае не надо готовить столбцы с запаздыванием, а можно обойтись одним столбцом Зарплата, хотя величину смещения придется занести в столбец G. В ячейку I2 введите формулу. В результате получите значения для функции автокорреляции, как показано на рис.2.4. Сравнивая эти значения, можно сделать вывод, что зарплата линейно зависит от предыдущего месяца, так как коэффициент r1 наибольший, следовательно, ряд содержит только тенденцию и не содержит циклических колебаний. Возможно, в уравнение модели следует ввести y_{t-1}.

	G	H	I
1	лаг	ФА	ФА с функцией СМЕЩ
2	1	=КОРРЕЛ(E3:E18;D3:D18)	=КОРРЕЛ(СМЕЩ($E$2:$E$18;G2;0;18-G2);СМЕЩ($E$2:$E$18;0;0;18-G2))
3	2	=КОРРЕЛ(E4:E18;C4:C18)	=КОРРЕЛ(СМЕЩ($E$2:$E$18;G3;0;18-G3);СМЕЩ($E$2:$E$18;0;0;18-G3))
4	3	=КОРРЕЛ(E5:E18;B5:B18)	=КОРРЕЛ(СМЕЩ($E$2:$E$18;G4;0;18-G4);СМЕЩ($E$2:$E$18;0;0;18-G4))
5	4		=КОРРЕЛ(СМЕЩ($E$2:$E$18;G5;0;18-G5);СМЕЩ($E$2:$E$18;0;0;18-G5))
6	5		=КОРРЕЛ(СМЕЩ($E$2:$E$18;G6;0;18-G6);СМЕЩ($E$2:$E$18;0;0;18-G6))
7	6		=КОРРЕЛ(СМЕЩ($E$2:$E$18;G7;0;18-G7);СМЕЩ($E$2:$E$18;0;0;18-G7))
8			

Рис.2.5. Использование функции СМЕЩ для вычисления ФА

2.4. Авторегрессионные модели (AR)

Модели, которые используют в качестве лаговых объясняющих переменных значения зависимой переменной (y_{t-k}) называются авторегрессионными. В примере 4 были найдены коэффициенты автокорреляции, которые свидетельствуют о сильной линейной зависимости между уровнями ряда со смещением в один, два лага.

Рассмотрим модель:

$$y=\alpha+\beta t+\gamma\, y_{t-1}+\varepsilon.$$

Такая модель называется **AR(1)**.

Для построения уравнения регрессии и оценки его коэффициентов удобно использовать пакет Анализ данных из меню Сервис приложения Excel. Воспользуемся данными примера 4 для поиска коэффициентов модели **AR(1)**.

На рис. 2.6 приведены все необходимые данные для оценки построенного уравнения регрессии:

$$\hat{y}_t = 3{,}515 + 0{,}006t + 0{,}391 y_{t-1}.$$

Коэффициент детерминации R^2=0,777 немного увеличился по сравнению со значением 0,7215 в линейной регрессии(рис1.5). все коэффициент AR(1) можно считать значимыми, так как P-значения не превышают 15%. Качество всего уравнения хорошее значимостьF=0. прогноз

Прогноз зарплаты для t=18:

$$\hat{y}_{18} = 3{,}515 + 0{,}006 \cdot 18 + 0{,}391 y_{17} = 5{,}942.$$

Месяц	Лаг1	Зарплата , доллар
4	29.7405	29.6405
5	29.6405	29.9125
6	29.9125	30.117
7	30.117	30.0705
8	30.0705	29.853
9	29.853	29.9535
10	29.9535	29.88
11	29.88	30.125
12	30.125	31.011
13	31.011	30.265
14	30.265	30.145
15	30.145	30.45
16	30.45	31.36
17	31.36	31.854
18	31.854	30.732
19	30.732	30.256
20	30.256	31.032

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0.5916316
R-квадрат	0.350028
Нормированный R-квадрат	0.3066965
Стандартная ошибка	0.5044108
Наблюдения	17

Дисперсионный анализ

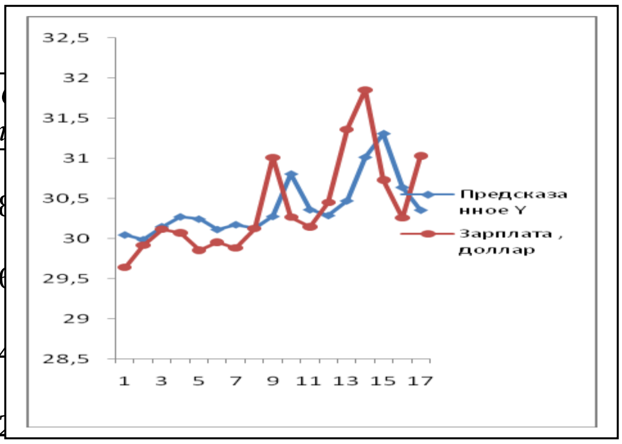
	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1	2.0552664	2.0552664	8.0779171	0.0123608
Остаток	15	3.8164537	0.2544302		
Итого	16	5.8717201			

	Коэффициенты	*Стандартная*	*t-статис*	*P-Знач*	*Нижние*	*Верхние*	*Нижние*	*Верхние*

		ошибка	тика	ение	95%	95%	95,0%	95,0%
Y-пересечение	12.325713	6.3575492	1.9387523	0.0715825	-1.2250824	25.876508	-1.2250824	25.876508
Переменная X 1	0.5959263	0.2096732	2.8421677	0.0123608	0.1490185	1.0428342	0.1490185	1.0428342

Рис. 2.6. Поиск коэффициентов модели AR(1)

ВЫВОД ОСТАТКА



Наблюдение	*Предсказанное Y*	*О... т...*
1	30.04886	0.408
2	29.989268	0.07
3	30.15136	0.034
4	30.273227	0.202
5	30.245516	-0.3925162
6	30.115902	-0.1624022
7	30.175793	-0.2957928
8	30.131992	-0.0069922
9	30.277994	0.7330058
10	30.805985	-0.5409849
11	30.361424	-0.2164238
12	30.289913	0.1600873
13	30.47167	0.8883298
14	31.013963	0.8400368
15	31.308351	-0.5763508
16	30.639721	-0.3837214
17	30.356061	0.6759395

Рис. 2.7. Значения зарплаты и графики временного ряда и приближения по модели AR(1)

Построим графики временного ряда с фактическими и теоретическими значениями уровней ряда. Для этого надо вычислить зарплату по построенному уравнению регрессии или воспользоваться предсказанными значениями пакета Анализ данных. Чтобы их получить включите опцию Остатки в диалоговом окне Регрессия.

МОДЕЛЬ AR(2) имеет вид: $y = \alpha + \beta t + \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \varepsilon$.

Построим эту модель для примера и сравним её с уже построенной моделью AR(1).

В качестве объясняющей переменной выберем ряд с лагом 1 и ряд с лагом 2. Для того, чтобы получить как можно больше сведений о новой модели воспользуйтесь пакетом Анализ данных.

Если сравнить нормированные коэффициенты детерминации для моделей AR(1) и AR(2), то можно сделать вывод, что модель AR(1) лучше описывает данные временного ряда, так как:

$$\overline{R}_1^2 = 0{,}743 > \overline{R}_2^2 = 0{,}715 \ .$$

Стандартная ошибка регрессии для первого уравнения меньше чем для второго. t-статистика для переменной лаг 2 приблизилась к единице, а ее P-значение увеличилось до 29%. Это означает, что коэффициент лаг2 незначим. Сравните самостоятельно два уравнения регрессии с помощью F-статистики.

Месяц	Лаг2	Лаг1	Зарплата , доллар
5	29.7405	29.6405	29.9125
6	29.6405	29.9125	30.117
7	29.9125	30.117	30.0705
8	30.117	30.0705	29.853
9	30.0705	29.853	29.9535
10	29.853	29.9535	29.88
11	29.9535	29.88	30.125
12	29.88	30.125	31.011
13	30.125	31.011	30.265
14	31.011	30.265	30.145
15	30.265	30.145	30.45
16	30.145	30.45	31.36
17	30.45	31.36	31.854
18	31.36	31.854	30.732
19	31.854	30.732	30.256
20	30.732	30.256	31.032

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0.8545217
R-квадрат	0.7302073
Нормированный R-квадрат	0.6627591
Стандартная ошибка	0.344291
Наблюдения	16

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	3	3.8498925	1.2832	10.82	0.0009
Остаток	12	1.4224358	0.1185		
Итого	15	5.2723282			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	42.1340	7.96591	5.2892	0.000	24.777	59.490	24.777	59.490
Месяц	0.13237	0.0318879	4.1511	0.001	0.0628	0.2018	0.0628	0.2018
Лаг2	-0.8022	0.2313214	-3.468	0.004	-1.306	-0.298	-1.3062	-0.2983
Лаг1	0.36153	0.2043485	1.7691	0.102	-0.083	0.8067	-0.0837	0.8067

Рис. 2.10. Данные для AR(2)

Можно сделать вывод, что модель AR(1) является предпочтительней модели AR(2).

Рассматривать авторегрессионные модели более высокого порядка не имеет смысла, так как коэффициенты автокорреляции уменьшаются с увеличением лага.

3. Моделирование сезонных и циклических колебаний

АДДИТИВНАЯ И МУЛЬТИПЛИКАТИВНАЯ МОДЕЛИ ВРЕМЕННОГО РЯДА

Существует несколько подходов к анализу структуры временных рядов, содержащих сезонные или циклические колебания.

Простейший подход — расчет значений сезонной компоненты методом скользящей средней и построение аддитивной или мультипликативной модели временного ряда. Общий вид аддитивной модели следующий:

$$Y = T + S + E. \tag{3.1}$$

Эта модель предполагает, что каждый уровень временного ряда может быть представлен как сумма трендовой (T), сезонной (S) и случайной (E) компонент. Общий вид мультипликативной модели выглядит так:

$$Y = T * S * E.$$

Эта модель предполагает, что каждый уровень временного ряда может быть представлен как произведение трендовой (T), сезонной (S) и случайной (E) компонент. выбор одной из двух моделей осуществляется на основе анализа структуры сезонных колебаний. Если амплитуда колебаний приблизительно постоянна , строят аддитивную модель временного ряда, в которой значения сезонной компоненты предполагаются постоянными для различных циклов. Если амплитуда сезонных колебаний возрастает или уменьшается, строят мультипликативную модель временного ряда, которая ставит уровни ряда в зависимость от значений сезонной компоненты.

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

Процесс построения модели включает в себя следующие шаги.

1. Выравнивание исходного ряда методом скользящей средней.
2. Расчет значений сезонной компоненты S.
3. Устранение сезонной компоненты из исходных уровней ряда и получение выравненных данных (T + E) в аддитивной или (T * E) в мультипликативной модели.
4. Аналитическое выравнивание уровней (T + E) или (T * E) и расчет значений T с использованием полученного уравнения тренда.
5. расчет полученных по модели значений (T + S) или (T * S).
6. Расчет абсолютных и/или относительных ошибок.

Если полученные значения ошибок не содержат автокорреляции, ими можно заменить исходные уровни ряда и в дальнейшем использовать временной ряд ошибок E для анализа взаимосвязи исходного ряда и других временных рядов.

Подробнее методику построения каждой из моделей рассмотрим на примерах.

Пример. Построение аддитивной модели временного ряда.

обратимся к данным об объеме потребления электроэнергии жителями района за последние четыре года, представленным в табл.1.

Шаг 1. Проведем выравнивание исходных уровней ряда методом скользящей средней. Для этого:

a) просуммируем уровнии ряда последовательно за каждые четыре квартала со сдвигом на один момент времени и определим условные годовые объемы потребления электроэнергии (гр. 3 табл.1)

b) разделив полученные суммы на 4, найдем скользящие средние (гр. 4 табл.1) Отметим, что полученные таким образом выравненные значения уже не содержат сезонной компоненты;

c) приведем эти значения в соответствие с фактическими моментами времени, для чего найдем средние значения из двух последовательных скользящих средних — центрированные скользящие средние(гр. 5 табл.1)

Расчет оценок сезонной компоненты в аддитивной модели Таблица 1

№ квартала, t	Потребление электроэнергии, Y_t	Итого за четыре квартала	Скользящая средняя за четыре квартала	Центрированная скользящая средняя	Оценка сезонной компоненты
1	2	3	4	5	6
1	9	—	—	—	—
2	8.3	—	—	—	—
3	7.6	34.1	8.525	8.45	-0.55
4	9.2	33.5	8.375	8.2875	-0.0125
5	8.4	32.8	8.2	8.35	0.75
6	7.6	34	8.5	8.3375	-0.8625

7	8.8	32.7	8.175	8.15	-0.25
8	7.9	32.5	8.125	8.3125	0.7125
9	8.2	34	8.5	8.2625	-0.5375
10	9.1	32.1	8.025	7.9375	0.0375
11	6.9	31.4	7.85	7.9125	-0.2875
12	7.2	31.9	7.975	7.65	-1.45
13	8.7	29.3	7.325	7.3125	0.4125
14	6.5	29.2	7.3	7.3625	0.1625
15	6.8	29.7	7.425	—	—
16	7.7	—	—	—	—

Шаг 2. Найдем оценки сезонной компоненты как разность между фактическими уровнями ряда и центрированными скользящими средними (гр. 6 табл 1). Используем эти оценки для разчета значений сезонной компоненты S (табл. 2). Для этого найдем средние за каждый квартал (по всем годам) оценки сезонной компоненты S_i. В моделях с сезонной компонентой обычно предполагается, что сезонные воздействия за период взаимопогашаются. В аддитивной модели это выражается в том, что сумма значений сезонной компоненты по всем кварталам должна быть равна нулю.

Расчет значений сезонной компоненты в аддитивной модели Таблица 2

Показатели	Год	№ квартала, i			
		I	II	III	IV
	1	—	—	-0.55	-0.0125
	2	0.75	-0.8625	-0.25	0.7125
	3	-0.5375	0.0375	-0.2875	-1.45
	4	0.4125	0.1625	—	—
Итого за i-й квартал (за все годы)		0.625	-0.6625	-1.0875	-0.75
Средняя оценка сезонной компоненты для i-го квартала, $\bar{S_i}$		0.2083333	-0.2208333	-0.3625	-0.25
Скорректированная сезонная компонента, Si		0.364583	-0.06458	-0.20625	-0.09375

Для данной модели имеем:

$$0,2083333 - 0,2208333 - 0,3625 - 0,25 = -0,625.$$

Определим корректирующий коэффициент:

$$K = -0,625 / 4 = -0,15625.$$

рассчитаем скорретированные значения сезонной компоненты как разность между ее средней оценкой и корректирующим коэффициентом K:

$$S_i = \bar{S_i} - K_i \qquad (3.2)$$

где i = 1:4.

Проверим условие равенства нулю суммы значений сезонной компоненты:

0,364583-0,06458-0,20625-0,09375 = 0.

Таким образом, получены следующие значения сезонной компоненты:

Ⅰ квартал: S_1 = 0,364583;

Ⅱ квартал: S_2 = -0,06458;

Ⅲ квартал: S_3 = -0,20625;

Ⅳ квартал: S_4 = -0,09375.

Занесем полученные значения в табл. 2. для соответствующих кварталов каждого года.

Шаг 3. Элиминируем влияние сезонной компонненты, вычитая ее значения из каждого уровня исходного временного ряда. Получим величины T + E = Y – S (гр. 4 табл. 3). Эти значения рассчитываются за каждый момент времени и содержат только тенденцию и случайную компоненту.

Расчет выравненных значений T и ошибок E в аддитивной модели Таблица 3

t	Y_t	S_i	$T+E=Y_t-S_i$	T	T+S	$E=Y_t$ - (T+S)	E^2
1	2	3	4	5	6	7	8
1	9	0.3645833	8.6354167	8.658088	9.022672	-0.02267	0.000514
2	8.3	-0.0645833	8.3645833	8.56951	8.504926	-0.20493	0.041995
3	7.6	-0.20625	7.80625	8.480931	8.274681	-0.67468	0.455195
4	9.2	-0.09375	9.29375	8.392353	8.298603	0.901397	0.812517
5	8.4	0.3645833	8.0354167	8.303775	8.668358	-0.26836	0.072016
6	7.6	-0.0645833	7.6645833	8.215196	8.150613	-0.55061	0.303174
7	8.8	-0.20625	9.00625	8.126618	7.920368	0.879632	0.773753
8	7.9	-0.09375	7.99375	8.038039	7.944289	-0.04429	0.001962
9	8.2	0.3645833	7.8354167	7.949461	8.314044	-0.11404	0.013006
10	9.1	-0.0645833	9.1645833	7.860882	7.796299	1.303701	1.699636
11	6.9	-0.20625	7.10625	7.772304	7.566054	-0.66605	0.443628
12	7.2	-0.09375	7.29375	7.683725	7.589975	-0.38998	0.152081
13	8.7	0.3645833	8.3354167	7.595147	7.95973	0.74027	0.547999
14	6.5	-0.0645833	6.5645833	7.506569	7.441985	-0.94199	0.887336
15	6.8	-0.20625	7.00625	7.41799	7.21174	-0.41174	0.16953
16	7.7	-0.09375	7.79375	7.329412	7.235662	0.464338	0.21561

Шаг 4. Определим компоненту T данной модели. Для этого проведем аналитическое выравнивание ряда (T + E) с помощью линейного тренда. Результаты аналитического выравнивания следующие:

Константа = 8,7466667

Коэффициент регрессии = -0,088578431

Таким образом, имеем следующий линейный тренд:

$$T = 8{,}7466667 - 0{,}088578431 * T$$

Подставляя в это уравнение значения t = 1, …, 16, найдем уровни T для каждого момента времени (гр. 5 табл. 3). График уравнения тренда приведен на рис. 1.

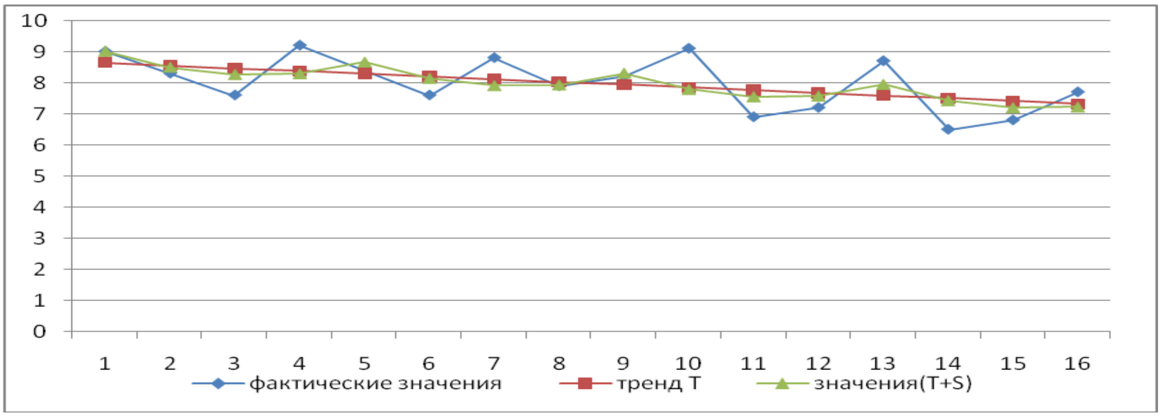


рис. 1. Потребление электроэнергии жителями района (фактические, выравненные и полученные по аддитивной модели значения уровней ряда)

Шаг 5. Найдем значения уровней ряда , полученные по аддитивой модели. Для этого прибавим к уровням T значения сезонной компоненты для соответствующих кварталов. Графически значения (T + S) представлены на рис. 1.

Шаг 6. В соответствии с методикой построения аддитивной модели расчет ошибки производится по формуле

$$E = Y - (T + S) \qquad (5.8)$$

Это абсолютная ошибка. Численные значения абсолютных ошибок приведены в гр. 7 табл. 3.

По аналогии с моделью регрессии для оценки качества построения модели или для выбора наилучшей модели можно применять сумму квадратов полученных абсолютных ошибок. Для данной аддитивной модели сумма квадратов абсолютных ошибок равна 1,10. По отношению к общей сумме квадратов отклонений уровней ряда от его среднего уровня, равной 71,59, эта величина составляет чуть более 1,5%:

$$(1 - 1{,}10 / 71{,}59) * 100 = 1{,}536$$

Следовательно, можно сказать, что аддитивная модель объясняет 98,5% общей вариации уровней временного ряда потребления электроэнергии за последние 16 кварталов.

ЗАКЛЮЧЕНИЕ

Различные данные, используемые для исследования явлений в макро- и микроэкономике, социологии и в других областях, выступают в форме временных (хронологических) рядов, которые часто еще называют рядами динамики. В частности, центральным моментом оценки и анализа конъюнктуры рынка является изучение тенденций и особенностей его развития, его устойчивости. Для выявления вектора и скорости развития рыночной ситуации используются динамические ряды показателей-индикаторов рынка.

Ряд динамики - последовательно расположенные в хронологическом порядке значения некоторого показателя. В своем изменении такой ряд отражает ход и направленность развития изучаемого явления. Элементами динамического ряда являются цифровые значения показателя, называемые уровнями ряда, и моменты или периоды времени, к которым относятся эти уровни. Если уровни ряда отражают состояние явления (процесса) на какой-либо момент времени (например, на начало месяца), то ряд именуют моментным; в противном случае интервальным (например, среднемесячные уровни продаж в течение года).

В зависимости от формы представления уровней ряд динамики может быть рядом показателей абсолютных, относительных или средних величин. Ряд динамики считается полным или с равноотстоящими уровнями, если одноименные моменты или периоды времени строго следуют один за другим. Если строгой хронологической последовательности нет, ряд называется неполным или с неравноотстоящими уровнями. Наконец, если уровни ряда есть одномерные величины, то и ряд динамики будет одномерным, иначе - многомерным.

С целью разностороннего изучения изменения явлений во времени ряды динамики подвергают разнообразной количественной обработке, используя при этом соответствующие разделы математики (метод средних, прогрессии, разложения, спектральный анализ). Простейшие вычислительные процедуры связаны с определением роста и прироста уровня динамического ряда, индекса сезонности и других характеристик. Для выявления тенденции развития социально-экономических явлений и взаимосвязи в динамике применяются специальные методы, например, метод скользящих средних, регрессионный анализ. Полученные на их основе модели можно использовать при составлении прогнозов, для выявления взаимосвязи между несколькими рядами динамики и выработки стратегии оптимального управления.

Раздел № 8. Динамические модели.

Под ожиданиями (expectations) понимаются предположения или мнения относительно будущих значений экономических переменных. Из-за неопределенности поведения рынка его участники формируют ожидания экономической конъюнктуры так, как это им необходимо для своей деятельности. Ошибочность или достоверность ожиданий очень важны для достижения равновесия в экономической системе. Если участники рынка принимают решения и действуют, исходя из неверных ожиданий, состояние экономики никогда не будет равновесным.

ГИПОТЕЗА РАЦИОНАЛЬНЫХ ОЖИДАНИЙ

Окончательная теория ожиданий еще не создана. Первой среди нескольких конкурирующих является гипотеза рациональных ожиданий (rational expectations hypothesis). Ее разработал в 1961 г. профессор Университета штата Индиана (США) Дж. Мут [1]. В основе гипотезы лежит принцип оптимизации приобретения и обработки информации с целью формирования мнения о будущем развитии конъюнктуры рынка.

Данный принцип предполагает, что индивиды не делают систематических ошибок в прогнозировании, напротив, их взгляды являются правильными. Игроки используют всю доступную информацию, составляя свое мнение о перспективах, или используют информацию до того момента, пока стоимость ее получения и обработки не превышает полученную выгоду. В крайней ситуации, когда имеется полная информация и отсутствует неопределенность, гипотеза рациональных ожиданий сводится к гипотезе совершенного поведения (perfect behavior hypothesis). Вариантом рациональных ожиданий являются т.н. частично рациональные ожидания (partly rational expectations), предполагающие, что агенты формируют свои ожидания долгосрочного равновесного состояния экономики, однако им неизвестна траектория, по которой система будет двигаться к равновесию. В связи с этим ожидания должны периодически пересматриваться с учетом отклонения фактических значений от прогнозируемых.

Гипотеза рациональных ожиданий приводит к двум обобщающим принципам, имеющим важное значение для финансового анализа:

1) Если меняется закон колебания наблюдаемой переменной, то и способ формирования ожиданий относительно нее тоже претерпевает изменения.

2) Ошибка прогноза в среднем равна нулю и не может быть предсказана заранее.

Кроме того, гипотеза рациональных ожиданий имеет еще одно очень важное следствие, затрагивающее концепцию эффективного рынка (efficient market concept). Если текущие цены на финансовые активы отражают всю доступную информацию, то возможность получения прибыли в принципе исключена.

Цены должны колебаться, т.е. изменяться непредсказуемо, в соответствии с законом случайного блуждания. На практике это может находить отражение в статических или неизменных ожиданиях (static expectations), когда предполагаемое значение переменной остается на прежнем уровне. Однако реалии финансовых рынков весьма далеки от такой идеализированной модели.

Существует, по крайней мере, две причины, по которым ожидания оказываются нерациональными. Во-первых, экономические агенты, хотя и располагают всей доступной информацией, используют ее не самым оптимальным образом. Во-вторых, игроки рынка могут упустить часть информации.

Поэтому даже самый лучший прогноз никогда не будет абсолютно точным.

ГИПОТЕЗА АДАПТИВНЫХ ОЖИДАНИЙ

Конкуренцию гипотезе рациональных ожиданий составляет гипотеза адаптивных ожиданий (adaptive expectations hypothesis). Адаптивные ожидания – это формирование будущих значений экономической переменной только на основе ее прошлых значений, на базе адаптации. Агенты корректируют свои прогнозы, исходя из последних наблюдений за рыночной конъюнктурой.

В гипотезе адаптивных ожиданий ключевую роль играет лаг ожиданий (expectations lag). Под ним понимается время пересмотра ожидаемого значения переменной в результате ее текущих изменений.

Гипотеза адаптивных ожиданий получила распространение в 50-60 годах XX века. Логика ее мало удовлетворительна, поэтому гипотезу используют в экономическом моделировании для упрощения анализа.

Есть несколько проблем, связанных с применением этой гипотезы. Одна из трудностей, возникающих при ее использовании, заключается в том, что невозможно определить информацию, на которой будет построен прогноз. Какова перспектива: пять, десять лет? Отсутствует также унификация в формировании ожиданий, и это делает применение гипотезы крайне проблематичным.

И, наконец, адаптивный подход имеет два спорных момента. Во-первых, гипотеза утверждает, что игроки не используют своего понимания экономических процессов, хотя на самом деле это не так. Во-вторых, существует бесчисленное множество механизмов и схем формирования адаптивных ожиданий, но нет способа выбрать из них один общепринятый при моделировании экономической системы.

АЛЬТЕРНАТИВНЫЕ ГИПОТЕЗЫ

Экономисты предпринимают регулярные попытки разработать альтернативные гипотезы ожиданий, но большинство из них терпят неудачи. Например, выдвигался поведенческий подход к ожиданиям (behavioural expectations approach), основывающийся на социальных и психологических факторах, с акцентом на межличностные контакты и массовую коммуникацию. Поведенческий подход с трудом поддается структуризации, его сложно смоделировать, что делает данный подход менее конкурентоспособным по сравнению с гипотезами рациональных и адаптивных ожиданий.

Экономистами предпринимались также попытки совместить рациональные и адаптивные ожидания, в результате чего появилась концепция ожиданий с распределенным лагом (distributed lag expectations). Она включает в спецификацию ожиданий не только текущее значение переменной, отражающее рациональность, но и ряд ее лаговых значений, учитывающих адаптивность.

Распределенный лаг требует, чтобы коэффициенты при лаговых значениях подчинялись определенной математической закономерности (или распределению) по времени – например, распределение может иметь форму перевернутой буквы U или убывать по экспоненте.

Разработаны специальные эконометрические методы для оценки особых форм уравнений с распределенным лагом. Возможен также вариант с частичной или неполной адаптацией (partly adaptive expectations), когда предыдущая информация учитывается не до конца. Типичным способом анализа уравнений с распределенным лагом бесконечной величины является преобразование Койка (Koyck transformation). Наиболее успешной альтернативной разработкой является гипотеза регрессивных или равновесных ожиданий (regressive expectations hypothesis). Она представляет собой попытку привязать ожидания к моделям валютного курса на базе фундаментальных макроэкономических факторов.

Регрессивные ожидания предполагают, что происходит неизбежный регресс, возвращение значения экономической переменной к своему равновесному уровню. С течением времени значение переменной может отклоняться в любую сторону от равновесия, но рано или поздно оно к нему возвращается. Такое допущение обычно используется в теориях, где валютный курс переоценен или недооценен, например, по отношению к паритетам покупательной способности или процентных ставок. Основные альтернативы гипотезе рациональных ожиданий могут быть объединены в группу экстраполятивных ожиданий (extrapolative expectations). Их обобщенное уравнение было выведено Дж. Френкелем и К. Фрутом [2]:

$E_t(S_{t+1}) = (1-\beta)S_t + \beta Z$ (1), где

S_t – валютный спот-курс в момент времени t;

E_t (...) – ожидания валютного курса в момент времени t;

S_{t+1} – валютный спот-курс в момент времени t+1;

β– коэффициент, характеризующий ожидания (betta)

Если Z= St-1, то речь идет об экстраполяции, распространении тенденций прошлого на будущий период. Можно различить три случая:

(1) betta < 0 – стадное поведение;

(2) betta = 0 – статические ожидания;

(3) betta > 0 – ожидания с распределенным лагом.

Уравнение (1) будет характеризовать адаптивные ожидания, если вместо Z подставить Et-1 (St). В случае, когда Z равно некоему Q, отражающему фундаментальный или равновесный валютный курс, то уравнение (1) будет описывать регрессивные ожидания.

ОБЗОР ПРИКЛАДНЫХ ИССЛЕДОВАНИЙ

Большинство специалистов по международной экономике считают гипотезу рациональных ожиданий главной концепцией ожиданий на валютном рынке. В целях эмпирического тестирования проверке подвергается непредубежденная или несмещенная гипотеза рациональности (unbiasedness hypothesis). Ее можно выразить с помощью следующего уравнения:

$S_{t+1} = \alpha + \beta E_t(S_{t+1}) + \varepsilon_{t,t+k}$, где

$\varepsilon_{t,t+k}$ – случайное колебание (white-noise error).

Несмещенная гипотеза рациональных ожиданий предполагает, что, во-первых, коэффициент betta должен быть равен единице, а, во-вторых, alfa должен равняться нулю.

Таким образом, ожидания валютного курса в момент времени t полностью оправдываются, и фактическое значение валютного курса соответствует его ожиданиям. Текущее значение валютного курса рассматривается как своеобразный дисконт ожидаемых фундаментальных макроэкономических факторов.

Эмпирические исследования изучают поведение как коэффициента betta, так и константы alfa. Одной из первых ввела в оценку ожиданий валютного курса регрессионно-корреляционный анализ К. Домингез [3]. Используя данные за 1983-1985 гг., она построила регрессионное уравнение на основе фактических данных и сравнила результаты расчетов с прогнозами. Выводы К. Домингез свидетельствуют, что несмещенная гипотеза рациональных ожиданий отвергается на однодневном, месячном и трехмесячном интервалах времени из-за того, что константа отлична от нуля.

Аналогичные выводы были получены большинством исследователей, изучавших поведение валютного курса в период 1985-1991 гг. [4]. Единственное исключение составляют исследования с использованием коинтеграции. Применение этой техники эконометрического анализа не позволяет полностью отвергнуть несмещенную гипотезу рациональных ожиданий. Например, в одной из новейших работ, подготовленной в Федеральном резервном банке Кливленда (США) в 2000 г. с использованием техники коинтеграции, подтверждается рациональное поведение игроков валютного рынка [5].

Несмещенная гипотеза также предполагает ортогональность, т.е. равенство alfa нулю. Иными словами, ожидания учитывают всю доступную информацию – неизвестного остатка, которым является alfa, не может быть. Абсолютно все исследователи, начиная с К. Домингез, находят, что экономические агенты не учитывают в своих прогнозах всю доступную информацию [6]. Наконец, эмпирические исследования обнаружили третью интересную закономерность в формировании ожиданий, т.н. искривленность (expectations twist). Оказалось, что долгосрочные и краткосрочные ожидания расходятся в противоположных направлениях. Краткосрочным ожиданиям свойственен эффект повального увлечения (bandwagon effects), тогда как долгосрочные ожидания достаточно стабильны.

Подводя короткие итоги по обзору исследований, проведенных в мире за последние двадцать лет, можно заметить следующее.

- Гипотеза рациональных ожиданий, по крайней мере, ее каноническая несмещенная версия, в целом отвергается в средне- и долгосрочном периоде. Однако использование техники коинтеграции не позволяет полностью отвергнуть гипотезу рациональных ожиданий.
- Основные альтернативы гипотезе рациональных ожиданий могут быть объединены в группу экстраполятивных ожиданий. Их тестирование выявило искривленность в ожиданиях: долгосрочные и краткосрочные ожидания расходятся в противоположных направлениях.
- Проверка гипотезы адаптивных ожиданий не дала четких результатов: выводы исследований зависят от горизонта прогноза и метода экстраполяции.
- Согласно регрессивным ожиданиям, в краткосрочном периоде валютный курс никогда не соответствует своему равновесному значению, но в длительной перспективе он близок к равновесию.
- В определенном приближении краткосрочные ожидания могут быть охарактеризованы как рациональные с признаками эффекта повального увлечения. В долгосрочном периоде ожидания успешно объясняются с помощью любой из экстраполятивных моделей: адаптивных, регрессивных ожиданий или ожиданий с распределенным лагом.

ТЕСТ ДЛЯ РОССИЙСКОГО РУБЛЯ

На Западе большое распространение получили обзоры частных прогнозов валютных курсов. Например, можно воспользоваться услугами Money Market Services, чьи данные представляют собой медианы прогнозов ведущих специалистов на недельную и месячную перспективу. Издается также Currency Forecasters' Digest, который покрывает статистику прогнозов валютных курсов 25 стран, включая развивающиеся рынки. Обзоры частных прогнозов позволяют оценить тип превалирующих на рынке ожиданий.

На отечественном рынке ничего подобного не существует и, видимо, еще долгое время не будет существовать. По этой причине единственное, что можно сделать, – решить обратную задачу: на основе фактических данных определить, какой тип ожиданий целесообразно использовать на российском валютном рынке.

Для решения этой задачи мной была обработана статистика ежедневных номинальных обменных курсов рубля к доллару США, которую ведет Банк России. Общий период наблюдения растянулся от января 1998 г. до сентября 2001 г. Пробным шаром для определения оптимальных ожиданий выступила регрессионная оценка уравнения (1). По результатам расчетов, приведенных в таблице 1, можно сделать следующие выводы.