ТЕМА 1
ВВЕДЕНИЕ. ЭКОНОМЕТРИКА И ЭКОНОМЕТРИЧЕСКОЕ МОДЕЛИРОВАНИЕ: ОСНОВНЫЕ ПОНЯТИЯ И ОПРЕДЕЛЕНИЯ

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

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

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

модели временных рядов,

— регрессионные модели с одним уравнением,

— системы одновременных уравнений.

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

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

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

Например, определение достоверных значений среднедушевого дохода непосредственно по результатам опросов и бухгалтерской отчётности может оказаться сложнее оценки изменения дохода на основе информации об изменении розничного оборота товаров и услуг, а также изменении общей суммы банковских вкладов населения.

Экзогенные переменные — это внешние по отношению к модели переменные, их значения определяются вне модели и поэтому они считаются фиксированными, обозначаются обычно как x.

Эндогенные переменные — это переменные, значения которых определяются внутри модели и обозначаются как y.

В эконометрике выделяют три типа данных:

I. ***Кросс секционные (перекрёстные)*** данные представляют ситуацию в группе переменных в отдельный момент времени. Таковыми, например, являются публикуемые в деловых разделах газет списки цен на различные акции, процентные ставки по разным видам вкладов и обменные курсы разных

валют. Другим примером может служить информация о продажах торговым предприятием в определённый день товаров различных групп (пищевых, хозяйственных и т.д.)

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

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

Эконометрическое моделирование состоит из следующих этапов:

1. *На постановочном этапе* формулируются конечные цели моделирования, определяется наборы возможных исследуемых (объясняемых) переменных $\overline{Y} = (y_1, y_2, \dots y_k)$ и факторных (объясняющих) переменных $\overline{X} = (x_1, x_2, \dots x_m)$.

2. *На предварительном этапе* осуществляется предварительный анализ экономической сути изучаемого явления, возможностей сбора и обработки статистических данных.

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

4. *Информационный этап* заключается в сборе информации (проведение наблюдений, использование материалов отчётности и т.д.) и предварительном анализе данных (проверка аномальных значений показателей, сглаживание, тестирование на наличие тенденции исследуемых показателей к изменению).

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

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

ТЕМА 2

ПАРНАЯ КОРРЕЛЯЦИЯ И РЕГРЕССИЯ

Изучение действительности показывает, что изменение каждого исследуемого (объясняемого) показателя находится в связи и взаимодействии с изменением объясняющих (факторных) показателей. Например, изменение производительности труда работников предприятия зависит от степени совершенства применяемого оборудования, технологии, организации труда, управления и других факторов.

Рассматривая зависимости между признаками, необходимо выделить два вида зависимостей: функциональные и корреляционные.

Функциональные связи характеризуются полным соответствием между изменением факторного признака (признаков) и исследуемого показателя. Так, величина начисленной зарплаты при повременной оплате труда однозначно определяется количеством отработанных часов.

В корреляционных связях между изменением факторного и результативного признаков нет однозначного соответствия, воздействие факторов проявляется лишь в среднем при многократном наблюдении фактических данных. Например, чем больше у человека заработная плата, тем больше он тратит денег на покупку одежды. Однако, точную величину таких расходов при определенной величине заработной платы назвать нельзя. Можно только определить среднюю величину расходов на одежду у людей с определённым размером заработной платы.

В отличие от жёсткости функциональной связи корреляционные связи характеризуются множеством причин и следствий и устанавливаются лишь тенденции изменения исследуемого признака при изменении факторного признака (признаков).

2.1 Ковариация. Выборочный коэффициент парной корреляции

Основная задача ***корреляционного анализа*** заключается в выявлении взаимосвязи между случайными переменными путём оценки коэффициентов корреляции и детерминации, а также проверки значимости полученных значений.

В *эконометрике* корреляционный анализ применяется для *отбора факторов*, оказывающих наибольшее влияние на исследуемый показатель и *оценки качества* построенных эконометрических *моделей*.

Мерой взаимосвязи между двумя переменными *v* и *w* является выборочная ковариация, вычисляемая по правилу:

$$Cov(v,w)=\frac{1}{n-1}\sum_{i=1}^{n}\left[(v_i-v_{cp})(w_i-w_{cp})\right],$$

где $v_i, w_i, i=1,\dots n$ — результаты наблюдений, n — число наблюдений,

$$v_{cp} = \frac{1}{n}\sum_{i=1}^{n} v_i \; ; \quad w_{cp} = \frac{1}{n}\sum_{i=1}^{n} w_i \; .$$

Обозначения переменных специально выбраны отличные от *x* и *y*, чтобы подчеркнуть возможность наличия связи между двумя любыми переменными, не обязательно являющимися объясняющей и объясняемой переменными.

Существенным недостатком ковариации является зависимость от единиц, в которых измеряются переменные *v* и *w*. Если мы одни и те же данные запишем с использованием различных единиц измерения, то получим различные значения ковариации. То есть любое ненулевое значение ковариации само по себе не позволяет сделать вывод о тесноте связи между переменными.

Поэтому для измерения силы связи между двумя переменными используется парный коэффициент корреляции. Парный коэффициент корреляции является показателем тесноты связи между переменными *v* и *w* лишь в случае линейной зависимости между этими переменными.

В практических расчётах обычно используется выборочный парный коэффициент парной корреляции, определяемый по имеющемуся набору фактических данных:

$$r(v,w) = \frac{\sum_{i=1}^{n}\left[(v_i - v_{cp})\cdot(w_i - w_{cp})\right]}{\sqrt{\sum_{i=1}^{n}(v_i - v_{cp})^2 \cdot \sum_{i=1}^{n}(w_i - w_{cp})^2}} = \frac{Cov(v,w)}{S_v \cdot S_w} \; , \qquad (2.1)$$

где $S_v^2 = \frac{1}{n-1}\sum_{i=1}^{n}(v_i - v_{cp})^2$, $S_w^2 = \frac{1}{n-1}\sum_{i=1}^{n}(w_i - w_{cp})^2$ — выборочные значения дисперсии переменных *v* и *w*.

> **Замечание.** Коэффициенты парной корреляции можно получить, используя программу Microsoft Excel:
> Excel 2010 / Данные / Анализ данных / Корреляция.

Парный коэффициент корреляции обладает следующими свойствами:

1) принимает значение в интервале [–1;1], то есть $\left|r(v,w)\right| \le 1$;

2) не зависит от выбора начала отсчёта и единицы измерения $r(av+b;cw+d) = r(v,w)$, где *a, b, c, d* — постоянные величины, причём *a* и *c* — положительны;

3) если $r(v,w) > 0$, то между переменными имеется *прямая* связь, то есть при возрастании (убывании) одной из них другая также возрастает (убывает);

4) если $r(v,w) < 0$, то связь является *обратной*, то есть при возрастании одной переменной другая убывает;

5) если $r(v,w) = \pm 1$, то между переменными имеется функциональная линейная зависимость, а если $r(v,w) = 0$, то линейная связь между пере-

менными отсутствует; соответственно, чем ближе модуль коэффициента парной корреляции к единице, тем теснее связь между переменными.

Заметим, что при отсутствии линейной связи между двумя переменными, между ними может существовать тесная связь другого вида.

Пример 1. В таблице приведены данные об объёмах продаж мороженого в магазине за день *y*, в зависимости от температуры воздуха в городе *x1* и процента торговой надбавки *x2*. Видно, что спрос быстро растёт при повышении температуры воздуха. При наступлении очень высоких температур, предприятие резко увеличивает наценку, поскольку оказывается не в состоянии физически удовлетворить резко возрастающий спрос и сдерживает его повышением цен.

Требуется определить наличие между переменными линейных корреляционных связей, сделать выводы об их тесноте и охарактеризовать как прямые или обратные.

Таблица 1

y, тыс. р.	x_1, °C	x_2, %
2	5	20
3,5	10	20
5	15	20
12	20	20
22	25	20
40	30	25
42	35	50

Решение. Используя программу Корреляция, получим матрицу коэффициентов парной корреляции между всеми имеющимися переменными.

	y	x_1	x_2
y	1		
x_1	0,9494	1	
x_2	0,7229	0,6901	1

Тогда выборочные коэффициенты парной корреляции:

$$r(y, x1) \approx 0,9494; \quad r(y, x2) \approx 0,7229; \quad r(x1, x2) \approx 0,6901.$$

Следовательно, мы можем сказать, что между переменными *y* (объём продаж) и x_1 (температура воздуха) имеется тесная прямая линейная связь. Между переменными x_1 (температура воздуха) и x_2 (торговая наценка) также наблюдается тесная прямая линейная зависимость. То же самое можно сказать о взаимосвязи между переменными *y* и x_2.

Для того чтобы проверить, можем ли мы делать вывод о наличии линейной корреляционной связи между переменными по полученному значению коэффициента парной корреляции производится оценка его значимости, то есть определяется действительно ли полученное значение отражает наличие линейной связи, или же ненулевое значение коэффициента получено в результате случайных колебаний показателей или является следствием погрешности в вычислениях.

2.2 Оценка значимости выборочного коэффициента парной корреляции

Для оценки значимости выборочного коэффициента парной корреляции применяется *t*-критерий Стьюдента. При этом фактическое значение этого критерия определяется по формуле:

$$t_{набл} = \sqrt{\frac{r^2}{1-r^2}(n-2)}\,, \qquad (2.2)$$

где *n* — число наблюдений.

Полученное значение сравнивается с табличным критическим значением $t_{кр}(\alpha,\nu)$, зависящим от уровня значимости α и числа степеней свободы $\nu = n-2$. Критическое значение может быть найдено по соответствующим таблицам, а при использовании табличного процессора Excel — с помощью функции СТЬЮДРАСПОБР (α ; ν).

При $t_{набл} > t_{кр}$ полученное значение коэффициента корреляции *r* признается значимым, то есть между переменными имеется линейная корреляционная зависимость.

Для рассмотренного **Примера 1** при $\alpha = 0{,}1$ с учётом количества степеней свободы $\nu = 7-2 = 5$ критическое значение $t_{кр}(0{,}1;5) = 2{,}02$. Вычислим $t_{набл}$ для каждой пары переменных и сделаем вывод о значимости соответствующих коэффициентов корреляции.

Для пары переменных *y, x1* :

$$t_{набл} = \sqrt{\frac{0{,}9494 \cdot 0{,}9494}{1-0{,}9494 \cdot 0{,}9494}} \cdot 5 \approx 6{,}76 \quad > \quad t_{кр} = 2{,}02\,.$$

Следовательно, значение коэффициента $r(y, x1) = 0{,}9494$ является значимым.

Для пары переменных *y,* x_2 :

$$t_{набл} = \sqrt{\frac{0,7229 \cdot 0,7229}{1 - 0,7229 \cdot 0,7229}} \cdot 5 \approx 2,34 \quad > \quad t_{кр} = 2,02 .$$

Следовательно, мы можем утверждать, что значение коэффициента $r(y, x2) = 0,7229$ является значимым.

Для пары переменных x_1, x_2 :

$$t_{набл} = \sqrt{\frac{0,6901 \cdot 0,6901}{1 - 0,6901 \cdot 0,6901}} \cdot 5 \approx 2,13 \quad > \quad t_{кр} = 2,02 .$$

Следовательно, значение коэффициента $r(x1, x2) = 0,6901$ является значимым.

Поскольку мы выбрали уровень значимости $\alpha = 0,1$, то с вероятностью 10% мы сделали ошибочные выводы, а с вероятностью $P = 1 - \alpha = 0,9$ наши выводы верны.

2.3 Модель парной регрессии. Основные понятия. Линейная парная регрессия

Регрессионное уравнение, разрешённое относительно исследуемой переменной *y* при наличии одной факторной переменной *x*, в общем виде записывается как:

$$y_p = f(x),$$

и показывает, каково будет *в среднем* значение переменной *y*, если переменная *x* примет конкретное значение. Индекс *p* указывает на то, что мы получаем расчётное значение переменной *y*. Мы говорим *в среднем*, поскольку под влиянием неучтённых в модели факторов и в результате погрешностей измерения фактическое значение переменной *y* может принимать различные значения для одного значения *x*.

Если *f(x)* является линейной функцией, то мы имеем общий вид модели парной линейной регрессии:

$$y_p = a + b \cdot x, \tag{2.3}$$

где *a* — постоянная величина (или свободный член уравнения), *b* — коэффициент регрессии, определяющий наклон линии, вдоль которой рассеяны наблюдения.

Коэффициент регрессии характеризует изменение переменной *y* при изменении значения *x* на единицу. Если $b > 0$, то переменные положительно коррелированы, если $b < 0$ — отрицательно коррелированы.

Фактическое значение исследуемой переменной y тогда может быть представлено в виде:

$$y = a + b \cdot x + \varepsilon \,, \qquad (2.4)$$

где ε — разность между фактическим значением (результатом наблюдения) и значением, рассчитанным по уравнению модели. Если модель адекватно описывает исследуемый процесс, то ε — независимая нормально распределённая случайная величина с нулевым математическим ожиданием ($M\varepsilon = 0$) и постоянной дисперсией ($D\varepsilon = \sigma^2$). Наличие случайной компоненты ε отражает тот факт, что присутствуют другие факторы, влияющие на исследуемую переменную и не учтённые в модели.

2.4 Определение параметров линейной парной модели методом МНК

Для оценки параметров a и b линейной парной регрессии с использованием имеющегося набора результатов наблюдений наиболее часто используют метод наименьших квадратов (МНК), который минимизирует сумму квадратов ε_i — отклонения результатов наблюдений y_i от рассчитанных по линейной модели (2.3) значений y_{pi}:

$$b = \frac{\sum\limits_{i=1}^{n} \left[(x_i - x_{cp}) \cdot (y_i - y_{cp}) \right]}{\sum\limits_{i=1}^{n} (x_i - x_{cp})^2}, \qquad a = y_{cp} - b \cdot x_{cp} \,. \qquad (2.5)$$

Такое решение может существовать только при выполнении условия $\sum\limits_{i=1}^{n} (x_i - x_{cp})^2 \neq 0$, то есть когда не все наблюдения проводились при одном и том же значении факторной переменной (сумма квадратов равна нулю, если каждое слагаемое равно нулю). Это условие называется ***условием идентифицируемости*** модели.

> **Замечание.** Для определения коэффициентов парной линейной регрессии можно использовать программу Microsoft Excel:
> Excel 2010 / Данные / Анализ данных / Регрессия.

По данным, приведённым в **Примере 1**, построим линейную модель для объёма продаж мороженного y в зависимости от температуры воздуха *x1*.

Используя программу Регрессия, в качестве «входного интервала X» выбираем значения фактора *x1*. Результаты вычислений представлены в таблицах:

ВЫВОД ИТОГОВ

Регрессионная статистика

Множественный R	0,949439
R-квадрат	0,901435
Нормированный R-квадрат	0,881722
Стандартная ошибка	5,868804
Наблюдения	7

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1	1575	1575	45,72791	0,001074
Остаток	5	172,2142857	34,4428571		
Итого	6	1747,214286			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	**–11,9286**	4,960044437	–2,4049324	0,061246	–24,6788	0,821629
x_1	**1,5**	0,221819931	6,76224177	0,001074	0,929794	2,070206

ВЫВОД ОСТАТКА

Наблюдение	*Предсказанное y*	*Остатки*
1	–4,428571429	6,428571429
2	3,071428571	0,428571429
3	10,57142857	–5,571428571
4	18,07142857	–6,071428571
5	25,57142857	–3,571428571
6	33,07142857	6,928571429
7	40,57142857	1,428571429

Коэффициенты модели содержатся в третьей таблице (столбец *Коэффициенты*). Таким образом уравнение парной линейной модели имеет вид:

$$y_p = -11,93 + 1,5 \cdot x1 .$$

Коэффициент регрессии b = 1,5 , следовательно, при увеличении температуры воздуха на 1 градус (x_1) объём продаж мороженого (y) увеличивается в среднем на 1,5 тысячи рублей.

Исходные данные наблюдений и результаты моделирования приведены на рис. 1.

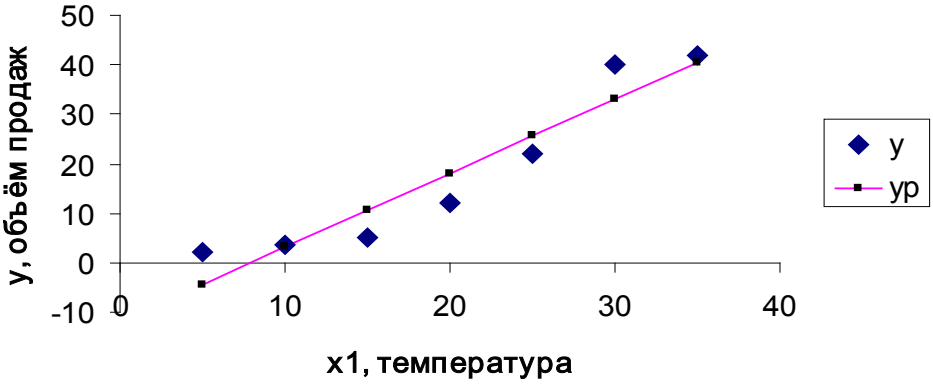


Рис 1 — Модель парной линейной регрессии

2.5 Проверка значимости параметров парной линейной модели

Поскольку в результате наблюдений мы имеем случайные значения y_i, то и вычисленные с их помощью параметры парной линейной модели a и b также являются случайными величинами. Для оценки надёжности полученных значений a и b производится проверка их значимости с использованием ***стандартной ошибки оценки,*** которая, в свою очередь, определяется по значениям ряда остатков ε_i:

$$S_{cm} = \sqrt{\frac{\sum_{i=1}^{n} \varepsilon_i^2}{n-m-1}}, \qquad (2.6)$$

где n — количество наблюдений, m — количество факторных переменных в модели. Выражение (2.6) для определения стандартной ошибки оценки будет использоваться нами в дальнейшем неоднократно, поскольку применимо в случае нелинейных моделей, а также при наличии в модели двух и более факторных переменных, то есть является универсальным.

Собственно проверка значимости параметров линейной модели производится в три этапа, аналогично тому, как это делалось для проверки значимости выборочного коэффициента корреляции.

На первом этапе вычисляются t-статистики:

$$t_a = \frac{a}{S_a}, \quad t_b = \frac{b}{S_b}, \tag{2.7}$$

где

$$S_a = S_{ст} \cdot \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - x_{ср})^2}}, \qquad S_b = S_{ст} \cdot \sqrt{\frac{1}{\sum_{i=1}^{n}(x_i - x_{ср})^2}}. \tag{2.8}$$

На втором этапе определяется критическое значение $t_{кр}(\alpha;\ n–m–1)$ по таблицам или с помощью функции СТЮДРАСПОБР в Excel. Уровень значимости α задаётся, а число степеней свободы вычисляется по числу наблюдений n и числу факторов m (в парной модели фактор x единственный).

Наконец, на третьем этапе абсолютные значения t-статистик сравниваются с критическими значениями $t_{кр}$. Если абсолютное значение больше табличного, то соответствующий параметр (коэффициент уравнения) считается значимым. В противном случае коэффициент значимым не является, то есть его можно положить равным нулю.

> **Замечание.** t-статистики для коэффициентов уравнения регрессии можно получить, используя программу Microsoft Excel: Excel 2010 / Данные / Анализ данных / Регрессия.

Произведём проверку значимости линейной модели парной регрессии, которую мы построили по данным **Примера 1.** Используя итоги программы Регрессия (3 таблица)

	Коэффициенты	*Стандартная ошибка*	***t*-статистика**	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	–11,9286	4,960044437	**–2,4049324**	0,061246	–24,6788	0,821629
x_1	1,5	0,221819931	**6,76224177**	0,001074	0,929794	2,070206

запишем значения t-статистики для коэффициентов уравнения регрессии:

$$t_a \approx -2,4;\ t_b \approx 6,76.$$

При уровне значимости 10% и числе степеней свободы 7–1–1 = 5 имеем $t_{кр}$=2,02. Поскольку абсолютные значения t-статистик для обоих параметров больше критического значения, то с вероятностью 90% можно утверждать, что оба параметра линейного уравнения a и b являются значимыми.

2.6 Проверка выполнения предпосылок МНК

Проверка выполнения предпосылок МНК выполняется на основе анализа остаточной компоненты ε. Ряд остатков ε должен удовлетворять ряду требований, а именно: равенство нулю математического ожидания, случайный характер отклонений от математического ожидания, отсутствие автокорреляции и неизменность дисперсии остатков при изменении факторной переменной, нормальный закон распределения. Рассмотрим способы проверки этих условий:

1. Проверка равенства математического ожидания уровней ряда остатков нулю осуществляется в ходе проверки соответствующей $H_0 : |\varepsilon| = 0$. С этой целью строится *t*-статистика

$$\hat{t} = \frac{|\bar{\varepsilon}|}{S_\Sigma}\sqrt{n}, \qquad S_\Sigma = \sqrt{\frac{\sum_{t=1}^{n}(\varepsilon_t - \bar{\varepsilon})^2}{n-1}}, \tag{2.9}$$

где $\bar{\varepsilon}$ — среднее арифметическое значение уровней ряда остатков ε_t, S_Σ — среднеквадратическое отклонение для этой последовательности, рассчитанное по формуле для малой выборки.

На уровне значимости α гипотеза отклоняется, если $\hat{t} > t_{\alpha,\nu}$, где $t_{\alpha,\nu}$ — критерий распределения Стьюдента с доверительной вероятностью (1–α) и $\nu = n-1$ степенями свободы.

2. Для проверки условия случайности возникновения отдельных отклонений от тренда часто используется *критерий поворотных точек*. Значение случайной переменной считается поворотной точкой, если оно одновременно больше (или одновременно меньше) значений предыдущего и последующего члена. Если остатки случайны, то поворотная точка приходится в среднем примерно на каждые 1,5 наблюдения.

Существует определённая зависимость между средней арифметической $\bar{p}$, дисперсией σ_p^2 количества поворотных точек в ряде остатков p и числом членов исходного ряда наблюдений n. С использованием этих зависимостей критерий случайности отклонений от тренда с доверительной вероятностью 0,95 можно представить в виде:

$$p > \left[\frac{2}{3}(n-2) - 1{,}96\sqrt{\frac{16n-29}{90}}\right], \tag{2.10}$$

где квадратные скобки означают, что от результата вычисления в правой части необходимо взять целую часть (не путать с процедурой округления!).

Если неравенство (2.10) не выполняется, то ряд остатков нельзя назвать случайным (то есть он содержит регулярную компоненту) и, следовательно, модель не является адекватной.

3. Наличие (отсутствие) автокорреляции в отклонениях ε_t фактических значений от модели роста проще всего проверить с помощью критерия Дарбина — Уотсона. С этой целью строится статистика Дарбина — Уотсона (*d*-статистика), в основе которой лежит расчётная формула

$$d = \frac{\sum_{t=1}^{n} (\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^{n} \varepsilon_t^2}. \qquad (2.11)$$

Для формулирования вывода о наличии (отсутствии) автокорреляции полученное значение необходимо сравнить с критическими значениями d_1 (нижнее) и d_2 (верхнее), которые определяются по специальным таблицам для трёх уровней значимости (α=0,01; α=0,025; α=0,05). При сравнении могут возникнуть следующие ситуации:

$d < d_1$ — остатки содержат автокорреляцию;

$d_1 < d < d_2$ — область неопределённости, когда нет оснований принять или отвергнуть гипотезу о существовании автокорреляции;

$d_2 < d < 2$ — ряд остатков некоррелирован.

Если *d* превышает 2, то это свидетельствует о наличии отрицательной корреляции. Перед входом в таблицу такие значения следует преобразовать по формуле $d' = 4 - d$.

Если установлено наличие автокорреляции остатков, нужно улучшить модель (изменить кривые роста, попытаться выделить дополнительные регулярные компоненты и т.п.). Если же ситуация оказалась неопределённой, применяют другие критерии. В частности можно воспользоваться первым коэффициентом автокорреляции:

$$r_1 = \frac{\left(\sum_{t=2}^{n} \varepsilon_t \varepsilon_{t-1}\right)}{\sum_{t=1}^{n} \varepsilon_t^2}. \qquad (2.12)$$

Для суждения о наличии или отсутствии автокорреляции в исследуемом ряду фактическое значение коэффициента автокорреляции (2.12) сопоставляется с табличным (критическим) $r_{1\,кр}$ для 5 %-го или 1 %-го уровня значимости (вероятность допустить ошибку при принятии гипотезы о независимости уровней ряда). Если $r_1 < r_{1\,кр}$, то гипотеза об отсутствии автокорреляции в ряду может быть принята. Когда же фактическое значение больше табличного, делают вывод о наличии автокорреляции во временном ряду.

4. Неизменность дисперсии остатков при изменении факторной переменной (исследование на гетероскедастичность) обычно проверяется с помощью трёх тестов, в которых делаются различные предположения о зависимости

между дисперсией случайной компоненты и факторной переменной: тест ранговой корреляции Спирмена, тест Голдфельда — Квандта и тест Глейзера.

При малом объёме выборки для оценки гетероскедастичности может использоваться метод Голдфельда — Квандта. Для проведения такого теста необходимо выполнить следующие шаги:

— упорядочить *n* наблюдений по мере возрастания переменной *x*;

— разделить совокупность наблюдений на две группы (соответственно с малыми и большими значениями фактора *x*) и построить по каждой из групп уравнение регрессии;

— определить остаточную сумму квадратов для первой регрессии $S_1 = \sum_{i=1}^{n_1} (y_i - y_{p1i})^2$ и второй регрессии $S_2 = \sum_{i=n_1+1}^{n} (y_i - y_{p2i})^2$;

— вычислить отношения $F_{набл} = S_2/S_1$ (или S_1/S_2). В числителе должна быть большая сумма квадратов;

— полученное отношение необходимо сравнить с $F_{кр}(\alpha, k_1, k_2)$, где $k_1 = n_1 - m$, $k_2 = n_2 - m$. Здесь n_1 и n_2 — количество наблюдений попавших в 1-ю и 2-ю группы. Если $F_{набл} > F_{кр}$, то гетероскедастичность имеет место, то есть условие о неизменности дисперсии при изменении факторной переменной не выполняется.

1. Соответствие ряда остатков нормальному закону распределения проверим с помощью *R/S* — критерия:

$$R/S = \frac{(\varepsilon_{\max} - \varepsilon_{\min})}{S_{\Sigma}}. \qquad (2.13)$$

Полученное значение проверяется на предмет попадания в интервал, границы которого являются табличными значениями, и зависят от уровня доверия α и количества наблюдений *n*.

Если все четыре пункта проверки 1–5 дают положительный результат, делается вывод о том, что выбранная трендовая модель является адекватной реальному ряду наблюдений. Только в этом случае её можно использовать для построения прогнозных оценок. В противном случае модель нужно улучшать.

2.7 Оценка качества уравнения регрессии

Для общей оценки качества построенной эконометрической определяются такие характеристики как коэффициент детерминации, индекс корреляции, средняя относительная ошибка аппроксимации, а также проверяется значимость уравнения регрессии с помощью *F*-критерия Фишера. Перечисленные характеристики являются достаточно универсальными и могут применяться как для линейных, так и для нелинейных моделей, а также моделей с двумя и более факторными переменными. Определяющее значение при вычислении всех перечисленных характеристик качества играет ряд остатков ε_i, который вычисляется путем вычитания из фактических (полученных по наблюдениям)

значений исследуемого признака y_i значений, рассчитанных по уравнению модели y_{pi}.

Коэффициент детерминации

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \varepsilon_i^2}{\sum_{i=1}^{n} \left(y_i - y_{cp} \right)^2} = \frac{\sum_{i=1}^{n} \left(y_{pi} - y_{cp} \right)^2}{\sum_{i=1}^{n} \left(y_i - y_{cp} \right)^2} \qquad (2.14)$$

показывает, какая доля изменения исследуемого признака учтена в модели. Другими словами коэффициент детерминации показывает, какая часть изменения исследуемой переменной может быть вычислена, исходя из изменений включённых в модель факторных переменных с помощью выбранного типа функции, связывающей факторные переменные и исследуемый признак в уравнении модели.

Коэффициент детерминации R^2 может принимать значения от 0 до 1. Чем ближе коэффициент детерминации R^2 к единице, тем лучше качество модели.

Замечание. Для определения коэффициента детерминации можно использовать программу Microsoft Excel:
Excel 2010 / Данные / Анализ данных / Регрессия.

Индекс корреляции можно легко вычислить, зная коэффициент детерминации:

$$R = \sqrt{R^2} \ . \qquad (2.15)$$

Индекс корреляции R характеризует тесноту выбранного при построении модели типа связи между учтёнными в модели факторами и исследуемой переменной. В случае линейной парной регрессии его значение по абсолютной величине совпадает с коэффициентом парной корреляции $r(x, y)$, который мы рассмотрели ранее, и характеризует тесноту линейной связи между x и y. Значения индекса корреляции, очевидно, также лежат в интервале от 0 до 1. Чем ближе величина R к единице, тем теснее выбранный вид функции связывает между собой факторные переменные и исследуемый признак, тем лучше качество модели.

Средняя относительная ошибка аппроксимации

$$E_{отн.ср} = 100 \cdot \frac{1}{n} \sum_{i=1}^{n} \left| \frac{\varepsilon_i}{y_i} \right| \qquad (2.16)$$

выражается в процентах и характеризует точность модели.

Достаточная точность модели при решении практических задач может определяться, исходя из соображений экономической целесообразности с

учётом конкретной ситуации. Широко применяется критерий, в соответствии с которым точность считается:

— неудовлетворительной, если $E_{отн.ср.} > 15\%$;

— удовлетворительной, если $E_{отн.ср} < 15\%$;

— модель имеет высокую точность, если $E_{отн.ср} < 5\%$.

Не рекомендуется применять для анализа и прогноза модели с неудовлетворительной точностью, то есть, когда $E_{отн.ср.}$ больше 15%.

F-критерий Фишера используется для оценки значимости уравнения регрессии. Расчётное значение F-критерия определяется из соотношения:

$$F = \frac{(n-m-1)\cdot R^2}{m\cdot(1-R^2)}. \qquad (2.17)$$

Критическое значение F-критерия определяется по таблицам при заданном уровне значимости α и степенях свободы $\nu_1 = m,\ \ \nu_2 = n-m-1$ (можно использовать функцию FРАСПОБР в Excel). Здесь, по-прежнему, m — число факторов, учтённых в модели, n — количество наблюдений. Если расчётное значение больше критического, то уравнение модели признаётся значимым. Чем больше расчётное значение F-критерия, тем лучше качество модели.

Замечание. Для определения значение F-критерия можно использовать программу Microsoft Excel:
Excel 2010 / Данные / Анализ данных / Регрессия.

Определим характеристики качества построенной нами линейной модели для **Примера 1**. Используем итоги программы Регрессия (1 таблица) для определения коэффициента детерминации.

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,949439
R-квадрат	**0,901435**
Нормированный R-квадрат	0,881722
Стандартная ошибка	5,868804
Наблюдения	7

Коэффициент детерминации $R^2 \approx 0,901$.

Следовательно, в рамках линейной модели изменение объёма продаж на 90,1% объясняется изменением температуры воздуха.

Индекс корреляции $R = \sqrt{0,85} \approx 0,95$.

Значение индекса корреляции в случае парной линейной модели как мы видим, действительно по модулю равно коэффициенту корреляции между соответствующими переменными (объём продаж и температура). Поскольку полученное значение достаточно близко к единице, то можно сделать вывод о наличии тесной линейной связи между исследуемой переменной (объём продаж) и факторной переменной(температура).

Используем итоги программы Регрессия (2 таблица) для определения расчётного значения *F*-критерия.

Дисперсионный анализ

	df	*SS*	*MS*	***F***	*Значимость F*
Регрессия	1	1575	1575	**45,72791**	0,001074
Остаток	5	172,2142857	34,4428571		
Итого	6	1747,214286			

F-критерий Фишера $F \approx 45,73$.

Критическое значение $F_{кр}$ при $\alpha = 0,1$; $\nu_1 = 1$; $\nu_2 = 7–1–1 = 5$ равно 4,06. Расчётное значение *F*-критерия больше табличного, следовательно, уравнение модели является значимым.

Используем итоги программы Регрессия (4 таблица) для определения средней относительной ошибки аппроксимации.

ВЫВОД ОСТАТКА

Наблюдение	*Предсказанное y*	***Остатки***
1	–4,428571429	**6,428571429**
2	3,071428571	**0,428571429**
3	10,57142857	**–5,571428571**
4	18,07142857	**–6,071428571**
5	25,57142857	**–3,571428571**
6	33,07142857	**6,928571429**
7	40,57142857	**1,428571429**

Рассмотрим остатки модели ε_i, содержащиеся в столбце Остатки. Вычислим относительные погрешности $\left|\frac{\varepsilon_i}{y_i}\right|$. Тогда

средняя относительная ошибка аппроксимации

$$E_{отн.ср.} = \frac{100}{7} \cdot \left(\frac{6,43}{2} + \frac{0,43}{3,5} + \ldots \frac{1,43}{42} \right) \approx 76,09 .$$

Построенная линейная модель парной регрессии имеет неудовлетворительную точность (>15%), и её не рекомендуется использовать для анализа и прогнозирования.

В итоге, несмотря на то, что большинство статистических характеристик удовлетворяют предъявляемым к ним критериям, линейная модель парной регрессии непригодна для прогнозирования объёма продаж в зависимости от температуры воздуха. Нелинейный характер зависимости между указанными переменными по данным наблюдений достаточно хорошо виден на рис. 1. Проведённый анализ это подтвердил.

2.8 Нелинейные модели парной регрессии

Если между исследуемыми и факторными переменными связь имеет нелинейный характер, то для построения модели необходимо использовать нелинейные функции.

Рассмотрим наиболее распространённые парные нелинейные модели.

Парабола второй степени определяет следующий вид модели:

$$y_p = a + bx + cx^2 . \qquad (2.18)$$

Параболическую модель целесообразно использовать, если связь меняет свой характер: прямая связь меняется на обратную или, наоборот, обратная связь меняется на прямую. Например, размер заработной платы работников физического труда в среднем растёт до некоторого возраста, а затем начинает убывать. Для определения параметров модели *a, b, c* модель (2.18) сводится путём замены переменных $x_1 = x; \quad x_2 = x^2$ к линейной модели двухфакторной модели

$$y_p = a + bx_1 + cx_2 \qquad (2.19)$$

Для оценки параметров модели вида (2.19), как будет показано далее, используется метод наименьших квадратов (МНК).

В основе ***гиперболической модели*** лежит уравнение гиперболы:

$$y_p = a + \frac{b}{x} \qquad (2.20)$$

Классическим примером гиперболической модели является *кривая Филипса,* характеризующая нелинейное соотношение между нормой безработицы *x* и процентом прироста заработной платы *y*: при росте *x* до некоторого уровня *y* также растёт, а при дальнейшем росте *x* рост *y* приостанавливается. Этот же характер связи проявляется при изучении зависимости расходов на единицу продукции сырья, материалов, топлива (то есть переменных затрат) от объёма

выпускаемой продукции. Другим примером гиперболической зависимости является зависимость времени оборота товаров в зависимости от величины товарооборота. Кривые Энгеля, описывающие долю доходов, расходуемых на непродовольственные товары, в зависимости от размера доходов также описываются гиперболическими функциями.

Сделав замену $x_1 = 1/x$, мы сведём уравнение (2.20) к линейному виду:

$$y_p = a + bx_1 \tag{2.21}$$

для оценки параметров которого используется МНК.

Степенная модель

$$y_p = a \cdot x^b \tag{2.22}$$

применяется для описания изменения спроса при изменении цены на товар. Параметр *b* в ней показывает, на сколько процентов уменьшится в среднем спрос, если цена увеличится на 1 % (то есть *b* — отрицательная величина) и называется *коэффициентом эластичности*. Логарифмирование соотношения (2.22) приводит его к линейному виду:

$$\ln y_p = \ln a + b \cdot \ln x\,. \tag{2.23}$$

Применение метода наименьших квадратов (с использованием прологарифмированных данных рядов наблюдений *x* и *y*) позволит нам найти коэффициенты уравнения (2.23) *ln a* и *b*, тем самым позволит найти параметры исходной степенной модели *a* и *b*.

В эконометрических исследованиях применяется также ***показательная модель***

$$y_p = a \cdot b^x\,. \tag{2.24}$$

Она также сводится к линейному виду путём логарифмирования:

$$\ln y_p = \ln a + \ln b \cdot x\,. \tag{2.25}$$

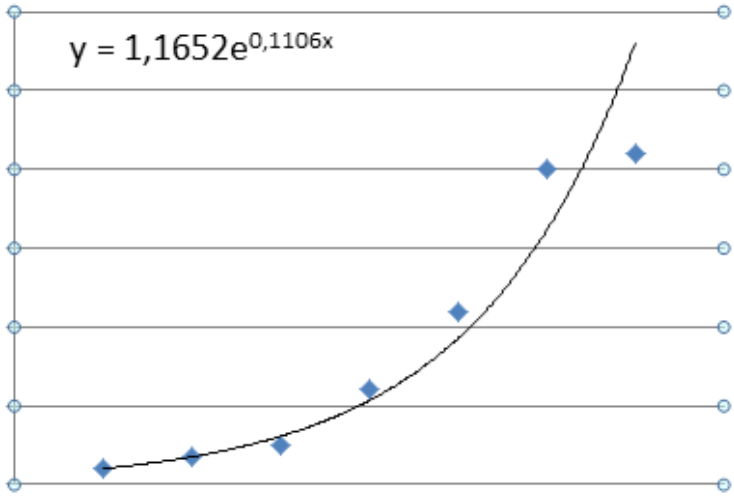
После логарифмирования ряда фактических значений *y* и применения МНК получим значения *ln a* и *ln b*. Возводя основание логарифма (в данном случае число *e*) в степень с использованием полученных значений, мы получим оценки параметров *a* и *b* исходной показательной модели.

Необходимо отметить, что не все нелинейные модели можно свести к линейной. Если модель не сводится к линейной, то она называется ***внутренне нелинейной***.

> **Замечание.** Для построения параболической, степенной, экспоненциальной моделей можно использовать инструменты программы Microsoft Excel:
> Excel 2010 / Вставка / Мастер диаграмм (точечная). Затем с помощью опций **Добавить линию тренда**…выбрать тип модели и показать уравнение

на диаграмме.

Построим ***показательную модель*** по данным **Примера 1**. Для этого построим поле корреляции исходных данных y и x_1, используя Мастер диаграмм. Добавим линию тренда — экспоненциальная модель.



Тогда $a \approx 1{,}165$; $b = e^{0{,}1106} \approx 1{,}117$. Зная параметры модели a и b, мы можем вычислить расчётные значения исследуемого признака по формуле (2.24) и составить ряд остатков.

Таблица 2

x_1	5,0	10,0	15,0	20,0	25,0	30,0	35,0
y	2,0	3,5	5,0	12,0	22,0	40,0	42,0
y_p	2,03	3,52	6,12	10,65	18,51	32,19	55,97
ε	–0,03	–0,02	–1,12	1,35	3,49	7,81	–13,97

Вычислим характеристики качества полученной показательной модели:

$$R^2 = 1 - \frac{0{,}03^2 + 0{,}02^2 + \ldots + 13{,}97^2}{16{,}07^2 + 14{,}57^2 + \ldots + 23{,}93^2} \approx 0{,}845; \qquad R = \sqrt{0{,}845} \approx 0{,}92;$$

$$F = \frac{5 \cdot 0{,}845}{(1 - 0{,}845)} \approx 27{,}2; \qquad E_{отн.ср.} = \frac{100}{7}\left(\frac{0{,}03}{2} + \frac{0{,}02}{3{,}5} + \ldots + \frac{13{,}97}{42}\right) \approx 14{,}9.$$

Характеристики качества показательной модели оказались лучше соответствующих характеристик линейной модели. Точность модели можно считать удовлетворительной.

Построив несколько моделей, выбрав из них лучшую, удовлетворяющую необходимым требованиям к качеству и точности модели, мы можем исполь-

зовать эту модель для прогнозирования.

2.9 Коэффициент эластичности для парных моделей регрессии

Коэффициент эластичности вычисляется по формуле:

$$Э = f'(x)\frac{x}{y},$$

где $f'(x)$ — первая производная. Для степенной модели $f'(x) = a \cdot b \cdot x^{b-1}$. Соответственно коэффициент эластичности окажется равным:

$$Э = a \cdot b \cdot x^{b-1} \cdot \frac{x}{a \cdot x^b} = \frac{a \cdot b \cdot x^b}{a \cdot x^b} = b.$$

Экономическое истолкование коэффициента эластичности следующее. Он показывает, на сколько процентов изменится в среднем результат, если фактор изменится на 1%. Коэффициент эластичности можно определить при наличии других видов связи, но только для степенной модели он представляет из себя постоянную величину. В других функциях коэффициент эластичности зависит от значений фактора *x*. В таблице приведены коэффициенты эластичности для некоторых математических функций.

Вид функции, Y		**Коэффициент эластичности, Э**
Линейная	$Yp = a + b \cdot x$	$Э = \frac{b \cdot x}{a + b \cdot x}$
Гиперболическая	$Y_p = a + \frac{b}{x}$	$Э = \frac{-b}{a \cdot x + b}$
Показательная	$Yp = a \cdot b^x$	$Э = x \cdot lnb$
Степенная	$Yp = a \cdot x^b$	$Э = b$

2.10. Прогнозирование с применением парного уравнения регрессии

Регрессионные модели могут использоваться для прогнозирования возможных ожидаемых значений исследуемой переменной при заданных (или определённых за рамками модели) значениях факторной переменной. При этом различают точечный и интервальный прогнозы.

Рассмотрим прогнозирование на основе ***парной линейной модели регрессии***

$$y_p = a_0 + a_1 \cdot x\,.$$

Точечный прогноз вычисляем путём подстановки в уравнение прогнозного значения факторной переменной:

Конец ознакомительного фрагмента.
Приобрести книгу можно
в интернет-магазине «Электронный универс»
(e-Univers.ru)