Специальная математика и основы статистики

Статистическое изучение взаимосвязи социально-экономических явлений

Вопрос 1. Корреляционный анализ

Исследование объективно существующих связей между социально-экономическими явлениями и процессами является важнейшей задачей статистики. В процессе исследования взаимосвязи вскрываются ***причинно-следственные отношения*** - такие отношения между явлениями и процессами, при которых изменение одного из исследуемых факторов - причины - ведет к изменению другого фактора - следствия. Эти отношения необходимо учитывать при регулировании и управлении, особенно в тех случаях, когда вы работаете со сложными системами, например, с производством, что бы правильно оценивать влияние управленческих решений на процесс его развития.

При исследовании социально-экономических явлений мы имеем дело с ***двумерными данными,*** когда каждая единица совокупности характеризуется информацией по двум показателям (например, размер заработной платы и уровень образования сотрудников компании), так как исследовать взаимосвязь можно только при совместном изучении обоих явлений.

Признаки по их сущности и значению для изучения взаимосвязи делятся па два класса. Признаки, обуславливающие изменения других, связанных с ними признаков, называются **факторными,** или просто факторами. Признаки, изменяющиеся под действием факторов, называются **результативными.**

В статистике различают *функциональную* и *корреляционную* связи. **Функциональной** называют связь, при которой определенному значению факторного признака соответствует одно и только одно значение результативного признака.

Если причинная зависимость проявляется не в каждом отдельном случае, а в общем, среднем, при большом числе наблюдений, то такая зависимость называется **корреляционной.** При **корреляционной связи** изменение среднего значения результативного признака обусловлено изменением признака-фактора.

По направлению выделяют связь прямую и обратную. **Прямая** - это связь, при которой с увеличением или с уменьшением значений факторного признака происходит увеличение или уменьшение значений результативного признака. Так, рост объемов производства способствует увеличению прибыли предприятия. В случае **обратной** связи значения результативного признака изменяются под воздействием факторного, но в противоположном направлении - с увеличением или с уменьшением значений факторного признака происходит уменьшение или увеличение значений признака-результата. Так, снижение себестоимости единицы производимой продукции вызывает рост прибыли.

Методы оценки связи между количественными переменными.

Корреляция – величина, отражающая наличие связи между явлениями, процессами и характеризующими их показателями.

Корреляционная зависимость – определение зависимости средней величины одного признака от изменения значения другого признака.

Формы проявления корреляционной связи между признаками:

1) *причинная* – зависимость результативного признака от вариации факторного признака;

2) *корреляционная связь между двумя следствиями общей причины* – здесь корреляцию нельзя интерпретировать как связь причины и следствия, поскольку оба признака – следствие одной общей причины;

3) *взаимосвязь признаков, каждый из которых и причина, и следствие* – каждый признак может выступать как в роли независимой переменной, так и в качестве зависимой переменной.

Задачи корреляционно-регрессионного анализа:

1) *выбор спецификации модели*, т. е. формулировки вида модели, исходя из соответствующей теории связи между переменными;

2) из всех факторов, влияющих на результативный признак, необходимо *выделить наиболее существенно влияющие факторы*;

3) парная регрессия достаточна, если имеется *доминирующий фактор*, который и используется в качестве объясняющей переменной;

4) исследовать, как *изменение одного признака меняет вариацию другого*.

Предпосылки корреляционно-регрессионного анализа:

1) *уравнение парной регрессии* характеризует связь между двумя переменными, которая проявляется как некоторая закономерность лишь в среднем в целом по совокупности наблюдений;

2) *в уравнении регрессии корреляционная связь* признаков представляется в виде функциональной связи, выраженной соответствующей математической функцией;

3) *случайная величина* ε включает влияние неучтенных в модели факторов, случайных ошибок и особенностей измерения;

4) *определенному значению признака-аргумента* отвечает некоторое распределение признака функции.

Недостатки корреляционно-регрессионного анализа:
1) невключение ряда объясняющих переменных:
* целенаправленный отказ от других факторов;
* невозможность определения, измерения определенных величин (психологические факторы);
* недостаточный профессионализм исследователя моделируемого;
2) агрегирование переменных (в результате агрегирования теряется часть информации);
3) неправильное определение структуры модели;
4) использование временной информации (изменив временной интервал, можно получить другие результаты регрессии);
5) ошибки спецификации:
* неправильный выбор той или иной математической функции;
* недоучет в уравнении регрессии какого-либо существенного фактора, (т.е. использование парной регрессии, вместо множественной);
6) ошибки выборки, так как исследователь чаще имеет дело с выборочными данными при установлении закономерной связи между признаками. Ошибки выборки возникают и в силу неоднородности данных в исходной статистической совокупности, что бывает при изучении экономических процессов;
7) ошибки измерения представляют наибольшую опасность. Если ошибки спецификации можно уменьшить, изменяя форму модели (вид математической формулы), а ошибки выборки – увеличивая объем исходных данных, то ошибки измерения сводят на нет все усилия по количественной оценке связи между признаками.

Корреляционные параметрические методы изучения связи

Корреляционные параметрические методы – методы оценки тесноты связи, основанные на использовании, как правило, оценок нормального распределения, применяются в тех случаях, когда изучаемая совокупность состоит из величин, которые подчиняются закону нормального распределения.

1) *Линейный коэффициент корреляции Пирсона* (r_{xy}) – количественная оценка и мера тесноты линейной связи между двумя переменными:

$$r_{xy} = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \sigma_y} = b \cdot \frac{\sigma_x}{\sigma_y}$$

где $r_{xy} \in [-1; 1]$. Если $r_{xy} = -1$, то наблюдается строгая отрицательная связь; $r_{xy} = 1$, то наблюдается строгая положительная связь; $r_{xy} = 0$, то линейная связь отсутствует. Если $r_{xy} < 0{,}3$, то связь слабая: $0{,}3 < r_{xy} < 0{,}7$ – средняя; $r_{xy} > 0{,}7$ – сильная, или тесная.

$cov(x,y)$ – *ковариация*, т.е. среднее произведение отклонений признаков от их средних квадратических отклонений.

Коэффициент корреляции может служить мерой зависимости случайных величин.

Проверка значимости проводится на основании критерия Стьюдента: если $t = r\sqrt{\frac{n-2}{1-r^2}} > t_{кр}(\alpha; k = n - m - 1)$, где n – объем выборки, m – количество факторов, то гипотеза о незначимости коэффициента корреляции отклоняется и переменные считаются зависимыми.

2) *Коэффициент детерминации* – квадрат линейного коэффициента корреляции, рассчитываемый для оценки качества подбора линейной функции: $R = \sqrt{1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}}$ при $R \in [0; 1]$. Чем ближе R к 1, тем теснее связь рассматриваемых признаков.

Проверка значимости проводится на основании критерия Фишера: если $F = \frac{s^2_{факт}}{s^2_{ост}} = \frac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\hat{y})^2} \cdot (n-2) > F_{кр}(\alpha; k_1 = m; k_2 = n - m - 1)$, где n – объем выборки, m – количество факторов в модели, то гипотеза о незначимости коэффициента детерминации отклоняется и выбранные переменные хорошо описывают линейное изменение y.

3) *Корреляция для нелинейной регрессии (индекс корреляции):* $R = \sqrt{\frac{1-\sigma^2_{ост}}{\sigma^2_y}}$

где σ^2_y – общая дисперсия результативного признака y, $\sigma^2_{ост}$ – остаточная дисперсия, определяемая исходя из уравнения регрессии: $y = f(x)$.

4) *Корреляция для множественной регрессии* оценивается с помощью показателя множественной корреляции и его квадрата – коэффициента детерминации. Показатель *множественной корреляции* характеризует тесноту связи рассматриваемого набора факторов с исследуемым признаком, или оценивает тесноту совместного влияния факторов на результат. Независимо от формы связи показатель множественной корреляции может быть найден как индекс множественной корреляции: $R_{yx_1 \dots x_n} = \sqrt{\frac{1-\sigma^2_{ост}}{\sigma^2_y}}$

где σ^2_y – общая дисперсия результативного признака;

$\sigma^2_{ост}$ – остаточная дисперсия для уравнения $y = f(x_1, \dots, x_n)$.

Корреляционные непараметрические методы изучения связи

Непараметрические методы не накладывают ограничений на закон распределения изучаемых величин. Их преимуществом является простота вычислений.

Непараметрические показатели связи позволяет судить о степени и тесноте связи не только, для количественных, но и для атрибутивных признаков:

1) *Коэффициент ассоциации*: $K = \frac{ad-bc}{ad+bc}$

2) *Коэффициент контингенции*: $K = \frac{ad-bc}{\sqrt{(a+b)(c+d)(d+b)(a+c)}}$

3) *Коэффициент корреляции рангов*: $K = 1 - \frac{6\sum d_i^2}{n^3 - n}$

Вопрос 2. Парная линейная регрессия: оценка параметров МНК-методом; оценка качества модели; оценка погрешности модели

Суть основной предпосылки построения эффективной эконометрической модели состоит в возможности разбиения Y на две части: объясненную и случайную: $Y = \hat{Y}(x) + \varepsilon$.

Объясненная часть случайной величины $\hat{Y}(x)$, формируется вариацией вектора независимых переменных x; ε – случайная составляющая (остаток).

Случайная величина ε называется также *возмущением*. Она порождается тремя источниками: спецификацией модели (т.е. либо влиянием не учтённых в модели факторов, либо неправильным выбором вида модели); выборочным характером исходных данных; особенностями измерения переменных.

Если случайная величина Y непрерывна, то объясненная часть $\hat{Y}(x)$ представляет собой некоторую неизвестную непрерывную функцию от регрессоров x_i: $\hat{Y}(x) = \varphi(x_1, \dots, x_n)$.

Естественной аппроксимацией (описанием) случайной функции φ(x) является оценка: $\varphi(x) \to M(X/x_1, \dots, x_n)$, где $M(X/x_1, \dots, x_n)$ – условное математическое ожидание, полученное при условии, что вектор независимых переменных принял конкретное (фиксированное) значение x_i.

Тогда *основную предпосылку построения эконометрической модели* можно записать так: $Y = M_x(Y) + \varepsilon$.

Уравнение: $\hat{Y}(x) = M_x(Y) = \varphi(x_1, \dots, x_n)$ называется *уравнением регрессии*. Заметим, что вид истинной функции φ(x) в последнем уравнении заранее неизвестен.

Замечание: Данная эконометрическая модель не всегда является регрессионной, т.е. объясненная часть случайной величины $\hat{Y}(x)$ не всегда равна своему условному математическому ожиданию: $\hat{Y}(x) \neq M_x(Y)$. Критерием того, что модель является регрессионной является условие $M_x(\varepsilon) = 0$.

Парная регрессия

Простая (парная) регрессия представляет собой регрессию между двумя переменными – y и x, т.е. модель вида: $y = f(x) + \varepsilon$, где y – зависимая переменная (результативный признак); x – независимая, или объясняющая, переменная (признак-фактор); ε – случайное возмущение.

Эконометрические модели с подобной спецификацией называются *парными регрессионными моделями*.

Наиболее часто на практике встречаются модели в виде $y = b_0 + b_1 \cdot x + \varepsilon$, которые называются *парными линейными моделями*.

В парной регрессии выбор вида математической функции f(x) может быть осуществлён тремя способами:

1) графическим;
2) аналитическим, т.е. исходя из теории изучаемого экономического процесса;
3) экспериментальным.

При изучении зависимости между признаками *графическим методом*, подбор вида уравнения регрессии основан на поле корреляции (рис. 1).

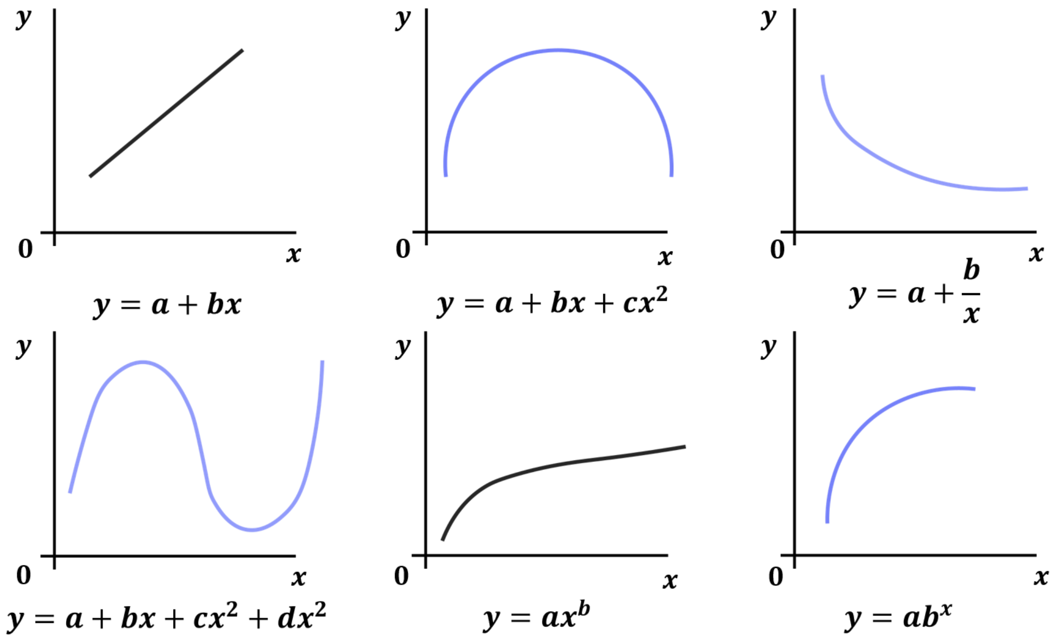


Рис. 1. Основные виды спецификаций

Аналитический способ выбора уравнения регрессии основан на изучении природы связи исследуемых признаков.

Пример 1. Изучается потребность предприятия в электроэнергии (y) в зависимости от объёма выпускаемой продукции (x).

Всё потребление электроэнергии можно разделить на две части: не связанное с производством (b_0) и связанное с объёмом выпускаемой продукции, пропорционально возрастающее с увеличением выпуска (b_1x).

Тогда зависимость потребления электроэнергии может быть выражена уравнением: $y = b_0 + b_1 \cdot x$.

Чаще всего в практических случаях выбор модели осуществляется *экспериментальным способом*, т.е. путём сравнения величины остаточной дисперсии $D_{\text{ост}}$, рассчитанной при разных моделях.
Если уравнение регрессии проходит через все точки корреляционного поля, то фактические значения результативного признака совпадают с теоретическими и $D_{\text{ост}} = 0$.
На практике имеет место некоторое рассеивание точек относительно линии регрессии, т.е. отклонение фактических данных от теоретических $(Y - \hat{Y})$ и $D_{\text{ост}} = \frac{1}{n}\sum(Y - \hat{Y})^2$. Предпочтение отдаётся модели, которая имеет наименьшую остаточную дисперсию.

Оценка параметров регрессионной модели
Параметры регрессионных моделей определяются при помощи статистических методов обработки наблюдений и, так как наблюдения – выборка ограниченного объёма, то полученные значения являются только *статистическими оценками* истинных значений параметров.
При подстановке статистических оценок параметров в уравнение регрессии получается *эмпирическая оценка* уравнения регрессии.
Оценки параметров модели, полученные по одной выборке, но использующие различные статистические методы, будут отличаться как по величине, так и по своим свойствам.
Как известно, точечные оценки параметров должны удовлетворять свойствам *несмещённости* (M(θ*)=θ), *состоятельности* ($\Theta^* \xrightarrow{P} \Theta$) и *эффективности* ($D_{\text{ост}}$ минимальна).
Метод наименьших квадратов (МНК) – метод оценивания параметров линейной регрессии, минимизирующий сумму квадратов отклонений наблюдений зависимой переменной от искомой линейной функции: $F\left(\widehat{b_0}, \widehat{b_1}\right) = \sum e_t^2 = \sum(y - \hat{y})^2 \to min.$
Решая данную задачу нахождения экстремума функции двух переменных, получаем следующие МНК-оценки: $\widehat{b_1} = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\overline{x^2} - \bar{x}^2}$; $\widehat{b_0} = \bar{y} - \widehat{b_1}\bar{x}$.
Экономический смысл параметров уравнения линейной парной регрессии:
- параметр b_1 показывает среднее изменение результата у при увеличении фактора х на единицу;
- параметр b_0 показывает уровень у, когда х = 0. Если х не может быть равен 0, то b_0 не имеет экономического смысла; в этом случае интерпретируется только знак при b_0: если $b_0 > 0$, то относительное изменение результата происходит медленнее, чем изменение фактора.
Пример 2. Оценить значения годовых доходностей акций компании А по значениям годовых доходностей акций компаний В (табл. 1).
Таблица 1

Доходность компаний за 10 периодов

t	Доходность компании А	Доходность компании В
1	-2,54	-5,31
2	26,50	16,84
3	4,44	0,07
4	17,12	10,03
5	10,19	4,98
6	13,88	7,52
7	4,55	0,23
8	10,28	5,53
9	11,76	5,94
10	11,89	6,09

Решение

Пусть y – доходность акций компании А, x – доходность акций компании В. Требуется построить регрессию $y = b_0 + b_1 \cdot x$. Сделаем вспомогательные расчёты (табл. 2).

Таблица 2

Дополнительные расчёты для вычисления МНК-коэффициентов

t	Доходность компании А, y	Доходность компании В, x	xy	x^2	y^2
1	-2,54	-5,31	13,4874	28,1961	6,4516
2	26,50	16,84	446,26	283,5856	702,25
3	4,44	0,07	0,3108	0,0049	19,7136
4	17,12	10,03	171,7136	100,6009	293,0944
5	10,19	4,98	50,7462	24,8004	103,8361
6	13,88	7,52	104,3776	56,5504	192,6544
7	4,55	0,23	1,0465	0,0529	20,7025
8	10,28	5,53	56,8484	30,5809	105,6784
9	11,76	5,94	69,8544	35,2836	138,2976
10	11,89	6,09	72,4101	37,0881	141,3721
Сумма	**108,07**	**51,92**	**987,055**	**596,7438**	**1724,0507**
Ср.зн.	**10,8**	**5,2**	**98,7**	**59,7**	**172,4**

Оценки параметров равны: $\hat{b}_1 = \dfrac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2} = \dfrac{98{,}7 - 10{,}8 \cdot 5{,}2}{59{,}7 - 5{,}2^2} = 1{,}3$

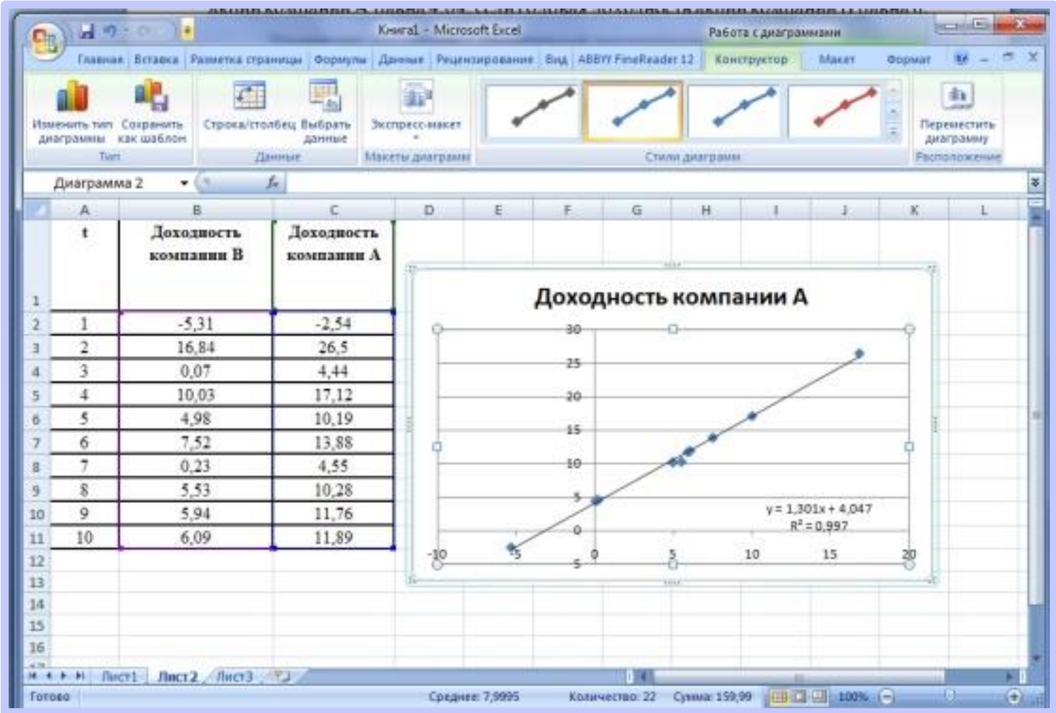
С экономической точки зрения b – показывает, что годовая доходность акций компании А увеличится на 1,3, если годовая доходность акций компаний В увеличится на 1.

$\hat{b}_0 = \overline{Y} - \hat{b}_1 \overline{X} = 10{,}8 - 1{,}3 \cdot 5{,}2 = 4{,}04$

С экономической точки зрения a – показывает, что годовая доходность акций компании А равна 4,04, если годовая доходность акций компаний В равна 0.

Уравнение регрессии с оценёнными параметрами имеет вид: $y = 4{,}04 + 1{,}3x$.

Проверим полученное уравнение в MS Excel:



Оценка качества модели

Для практического использования моделей регрессии большое значение имеет их *адекватность,* т.е. соответствие фактическим статистическим данным.

Анализ качества эмпирического уравнения линейной регрессии начинают с построения эмпирического уравнения регрессии, которое затем проверяют по следующей устоявшейся схеме:

* проверка статистической значимости коэффициентов уравнения регрессии;

* проверка общего качества уравнения регрессии;

* проверка свойств данных, выполнимость которых предполагалась при оценивании уравнения (проверка выполнимости предпосылок МНК).

Прежде, чем проводить анализ качества уравнения регрессии, необходимо определить дисперсии и стандартные ошибки коэффициентов, а также интервальные оценки коэффициентов.

Для построения доверительных интервалов необходимо вычислить:

$$s_{b_1} = \sqrt{\frac{\frac{\sum (y-\hat{y})^2}{n-2}}{\sum (x-\bar{x})^2}}, \quad s_{b_0} = \sqrt{\frac{\sum (y-\hat{y})^2}{n-2} \frac{\sum x^2}{n\sum (x-\bar{x})^2}},$$

$$t_{b_1} = \frac{b_1}{s_{b_1}}, \quad t_{b_0} = \frac{b_0}{s_{b_0}}.$$

Тогда границы доверительных интервалов:
$\Delta b_0 = t_{табл} s_{b_0}$; $\Delta b_1 = t_{табл} s_{b_1}$;
$b_0 - \Delta b_0 < \gamma_{b_0} < b_0 + \Delta b_0$; $b_1 - \Delta b_1 < \gamma_{b_1} < b_1 + \Delta b_1$.

1) *Проверка значимости (существенности) каждого коэффициента регрессии* осуществляется с помощью *t-критерия Стьюдента*.

Если $t_{b_i} = \frac{b_i}{s_{b_i}} > t_{кр}(\alpha; k = n - m - 1)$, то параметр b_i признается значимым (существенным). В этом случае, практически невероятно, что найденные значения параметров обусловлены только случайными совпадениями.

Следует отметить, что в социально-экономических исследованиях уровень значимости α обычно принимают равным 0,05.

2) *Проверка значимости (качества) уравнения регрессии* – значит установить, соответствует ли математическая модель, выражающая зависимость между переменными, экспериментальным данным; достаточно ли включенных в уравнение объясняющих переменных для описания зависимой переменной.

Проверка адекватности уравнения регрессии (модели) осуществляется с помощью *средней ошибки аппроксимации*, величина которой не должна превышать 12-15% (максимально допустимое значение): $\bar{A} = \frac{1}{n}\sum\left|\frac{y-\hat{y}}{y}\right| \cdot 100\%$

Оценка значимости уравнения регрессии в целом производится на основе *F-критерия Фишера*.

Если $F = \frac{s^2_{факт}}{s^2_{ост}} = \frac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\hat{y})^2} \cdot (n - m - 1) > F_{кр}(\alpha; k_1 = m; k_2 = n - m - 1)$, то признается статистическая значимость уравнения в целом (n – число наблюдений, m – число факторов в модели).

Характеристикой качества уравнения регрессии или соответствующей модели связи является *коэффициент детерминации*, который имеет вид: $R^2 = 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}$.

Коэффициент детерминации R^2 принимает значения в диапазоне от нуля до единицы $0 \le R^2 \le 1$ и показывает, какая часть дисперсии результативного признака (y) объяснена уравнением регрессии. Таким образом, значение R^2 является индикатором степени подгонки модели к данным (значение R^2 близкое к 1.0 показывает, что модель объясняет почти всю изменчивость соответствующих переменных).

Чтобы определить, при каких значениях R^2 уравнение регрессии следует считать статистически незначимым, что, в свою очередь, делает необоснованным его использование в анализе, рассчитывается тот же F-критерий Фишера. Связь F-критерия и коэффициента детерминации R^2 выражается формулой: $F = \frac{R^2}{1-R^2} \cdot (n - m - 1)$.

При анализе адекватности уравнения регрессии (модели) исследуемому процессу, возможны следующие варианты:

1. Построенная модель на основе F-критерия Фишера в целом адекватна и все коэффициенты регрессии значимы. Такая модель может быть использована для принятия решений и осуществления прогнозов.
2. Модель по F-критерию Фишера адекватна, но часть коэффициентов не значима. Модель пригодна для принятия некоторых решений, но не для прогнозов.
3. Модель по F-критерию адекватна, но все коэффициенты регрессии не значимы. Модель полностью считается неадекватной. На ее основе не принимаются решения и не осуществляются прогнозы.

Пример 2. (продолжение) Проверим значимость параметров и модели в целом.

Оценим: $r_{xy} = \frac{\overline{xy}-\bar{x}\cdot\bar{y}}{\sigma_x\sigma_y} = \frac{98{,}7-10{,}8\cdot5{,}2}{\sqrt{59{,}7-5{,}2^2}\cdot\sqrt{172{,}4-10{,}8^2}} = 0{,}995$

Проверим гипотезы H0:b_0=b_1=r=0 ($t_{кр}(0{,}05;8)=2{,}160$):

$$s_{b_1} = \sqrt{\frac{\sum(y-\hat{y})^2/(n-2)}{\sum(x-\bar{x})^2}} = \sqrt{\frac{1{,}583/8}{327{,}178}} = 0{,}0246;\ t_{b_1} = \frac{b_1}{s_{b_1}} = \frac{1{,}3}{0{,}0246} = 52{,}85$$

Так как, $t_{b_1} = 52{,}85 > t_{кр} = 2{,}16$, то гипотеза H0:$b_1$=0 отклоняется и b_1 признаётся значимым.

$$s_{b_0} = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}\cdot\frac{\sum x^2}{n\sum(x-\bar{x})^2}} = \sqrt{\frac{1{,}583}{8}\cdot\frac{596{,}74}{10\cdot327{,}17}} = 0{,}19;\ t_{b_0} = \frac{4{,}04}{0{,}19} = 21{,}7$$

Таким образом, гипотеза H0:b_0=0 отклоняется и b_0 признаётся значимым.

$$s_r = \sqrt{\frac{1-r^2}{n-2}} = \sqrt{\frac{1-0{,}995^2}{8}} = 0{,}035;\ t_r = \frac{0{,}995}{0{,}035} = 0{,}035$$

$$s_r = \sqrt{\frac{1-r^2}{n-2}} = \sqrt{\frac{1-0{,}995^2}{8}} = 0{,}035;\ t_r = \frac{r}{s_r} = \frac{0{,}995}{0{,}035} = 0{,}035$$

Таким образом, гипотеза H0:r=0 принимается и r признаётся незначимым.

Рассчитаем доверительные интервалы: $\Delta b_0 = t_{кр}s_{b_0} = 2{,}16\cdot0{,}19 = 0{,}4104$, тогда $b_0 \pm \Delta b_0 = 4{,}04 \pm 0{,}41$ и $3{,}63 < \gamma_{b_0} < 4{,}45$.

Доверительный интервал для b_0 показывает, что с вероятностью 95%, годовая доходность акций компании А будет варьироваться в пределах от 3,63 до 4,45, если годовая доходность акций компаний В равна 0.

$\Delta b_1 = 2{,}16\cdot0{,}0246 = 0{,}053$, тогда $b_1 \pm \Delta b_1 = 1{,}3 \pm 0{,}053$ и $1{,}247 < \gamma_{b_1} < 1{,}353$.

Доверительный интервал для b_1 показывает, что с вероятностью 95%, годовая доходность акций компании А вырастет в пределах от 1,247 до 1,353, если годовая доходность акций компаний В вырастет на 1.

Кроме того, доверительные интервалы позволяют построить оптимистический и пессимистический прогнозы развития процесса:

- пессимистический прогноз: $y=3{,}63+1{,}247x+\varepsilon$;
- оптимистический прогноз: $y=4{,}45+1{,}353x+\varepsilon$.

Проверим значимость уравнения ($F_{кр}(0{,}05;1;8)=4{,}67$). Сделаем дополнительные расчёты (табл. 3).

Таблица 3

Дополнительные расчёты для проверки гипотезы о значимости уравнения регрессии

t	Доход. комп. А, y	Доход. комп. В, x	y^	$(y-\hat{y})^2$	$(x-x_{ср})^2$	$(\hat{y}-y_{ср})^2$	$(y-y_{ср})^2$
1	-2,54	-5,31	-2,863	0,104	110,46	186,87	178,14
2	26,50	16,84	25,932	0,323	135,49	227,5	246,27
3	4,44	0,07	4,131	0,095	26,32	44,57	40,54
4	17,12	10,03	17,079	0,002	23,33	39,34	39,85
5	10,19	4,98	10,514	0,105	0,048	0,09	0,38
6	13,88	7,52	13,816	0,004	5,38	9,05	9,44
7	4,55	0,23	4,339	0,045	24,7	41,84	39,15
8	10,28	5,53	11,229	0,9	0,11	0,18	0,28
9	11,76	5,94	11,762	0,000	0,55	0,91	0,91
10	11,89	6,09	11,957	0,005	0,79	1,32	1,17
Сумма	**108,07**	**51,92**		**1,583**	**372,178**	**552,93**	**556,14**
Ср.зн.	**10,8**	**5,2**					

Тогда $F=\frac{s^2_{факт}}{s^2_{ост}}=\frac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\hat{y})^2}\cdot(n-2)=\frac{552{,}93}{1{,}583}\cdot 8=2794{,}33.$

Так как $F=2794{,}33>F_{кр}=4{,}67$, то гипотеза $H0{:}b_1=0$ отклоняется и уравнение регрессии признаётся значимым.

Коэффициент детерминации равен $R^2=1-\frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}=1-\frac{1{,}583}{556{,}14}=0{,}997$. Полученное значение говорит о том, что доходность акций компании А на 97% определяется доходностью акций компании В. Данный факт может свидетельствовать о том, что компания А является дочерней компанией (филиалом, структурным подразделением и т.п.) для компании В.

■ ■ ■

Вопрос 3. Нелинейная парная регрессия

Большинство экономических процессов лучше описываются нелинейными соотношениями. Однако, на практике, из всех существующих нелинейных функций применяются лишь те, которые могут с помощью замены переменных быть приведены к линейному виду. Данный процесс называется *линеаризация нелинейной модели*.

В моделях линейной регрессии (как парной, так и множественной) переменные имеют 1-ю степень (*модель, линейная по переменным*), а параметры выступают в виде коэффициентов при этих переменных (*модель, линейная по параметрам*).

Поэтому уравнения в нелинейных моделях могут быть нелинейными как по переменным, так и по параметрам.

1. Модели, нелинейные по переменным.

А) полиномиальные модели

В данных моделях, регрессоры, имеющие степень отличную от 1-ой заменяются другими независимыми переменными 1-ой степени, и к новой системе применяется обычный МНК. После оценки параметров новые переменные заменяются на первоначальные.

Б) гиперболическая регрессия

Для преобразования гиперболической регрессии $y = b_0 + b_1 \cdot \frac{1}{x} + \varepsilon$ к линейному виду используется замена $\tilde{x} = \frac{1}{x}$. Экономическая интерпретация параметров: b_0 – уровень эндогенной переменной при больших значениях регрессора; b_1 – скорость приближения к данному уровню.

Для преобразования регрессии $y = \frac{1}{\beta_0 + \beta_1 x + \varepsilon}$ к линейному виду используется замена $\tilde{y} = \frac{1}{y}$.

Для линеаризации $y = \frac{x}{\beta_0 + \beta_1 x + x \cdot \varepsilon}$ необходима замена как для регрессора $\tilde{x} = \frac{1}{x}$, так и для эндогенной переменной $\tilde{y} = \frac{1}{y}$.

2) Модели, нелинейные по параметрам

А) логарифмические модели

Рассмотрим уравнение вида $y = b_0 x^{b_1}$. Параметр b_0 – значение эндогенной переменной при равенстве регрессора единице; b_1 – эластичность переменной y по переменной x.

Если прологарифмировать обе части данного уравнения $\ln(y) = \ln(b_0) + \ln(b_1) \cdot \ln(x) = \alpha + \beta \cdot \ln(x)$, то получаем линейную относительно логарифмов функцию, которая называется *двойственной логарифмической моделью* или *моделью с постоянной эластичностью*, и вводя замену $\tilde{y} = \ln(y)$, $\tilde{x} = \ln(x)$ получаем линейную спецификацию, к которой применим МНК-метод.

Спецификации линейных моделей, включающие либо только логарифмы значений эндогенных переменных либо только логарифмы регрессоров, называются *полулогарифмическими*.

Модель со спецификацией $\ln(y) = \alpha + \beta \cdot x + \varepsilon$ называется *лог-линейной моделью*, для приведения которой к линейному виду

используется замена $\tilde{y} = \ln(y)$. В данной модели β – темп прироста переменной y по переменной x.
Модель со спецификацией $y = \alpha + \beta \cdot \ln(x) + \varepsilon$ называется *линейно-логарифмической моделью*, для приведения которой к линейному виду используется замена $\tilde{x} = \ln(x)$.

Вопрос 4. Множественная линейная и нелинейная регрессии

Множественная регрессия – регрессия между переменными y и $x_1, \ldots, x_m$, т. е. модель вида: $y = f(x_1, \ldots, x_m) + \varepsilon$,
где y – зависимая переменная (результативный признак); $x_1, \ldots, x_m$ – независимые, объясняющие переменные (признак-факторы); ε – случайное возмущение, или стохастическая переменная, включающая влияние неучтенных факторов в модели.
Основные типы функций, используемые при количественной оценке связей:
1) линейная функция: $y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_mx_m + \varepsilon$
Параметры $b_1, b_2, \ldots, b_m$ называются *коэффициентами регрессии* и характеризуют среднее изменение результата с изменением соответствующего фактора на единицу при неизменном значении других факторов, закрепленных на среднем уровне;
2) нелинейные функции:
- полиномиальные функции: $y = b_0 + b_1x + b_2x^2 + b_3x^3 + \ldots + b_mx^m + \varepsilon$
Параметры такой спецификации имеют конкретную интерпретацию: b_0 – значение эндогенной переменной при значении регрессора равном нулю, т.е. начальный уровень; b_1 – прирост эндогенной переменной при изменении регрессора на единицу (скорость роста); b_2 – скорость изменения скорости (ускорение роста); b_3 – изменение ускорения и т.д.
- степенные функции: $y = ax_1^{b_1}x_2^{b_2}\ldots x_m^{b_m} \cdot \varepsilon$
Параметры b1, b2,…, b_m – коэффициенты эластичности; показывают, на сколько процентов изменится в среднем результат при изменении соответствующего фактора на 1% при неизменности действия других факторов;
- гиперболические функции: $y = \dfrac{1}{a_0 + a_1x_1 + \ldots + a_mx_m + \varepsilon}$;
$y = \dfrac{x}{a_0 + a_1x_1 + \ldots + a_mx_m + \varepsilon}$; $y = a_0 + a_1\dfrac{1}{x_1} + \ldots + a_m\dfrac{1}{x_m} + \varepsilon$
- показательные функции, в частности экспоненциальные: $y = e^{a_0 + a_1x_1 + \ldots + a_mx_m + \varepsilon}$
- логарифмические функции, которые делятся на: двойственно-логарифмические модели $\ln y = a_0 + a_1 \ln x_1 + \ldots + a_m \ln x_m + \varepsilon$; лог-

линейные модели $\ln y = a_0 + a_1 x_1 + ... + a_m x_m + \varepsilon$; линейно-логарифмические модели $y = a_0 + a_1 \ln x_1 + ... + a_m \ln x_m + \varepsilon$.