Министерство образования и науки Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования
«Томский государственный архитектурно-строительный университет»

ВВЕДЕНИЕ В СТАТИСТИЧЕСКИЕ ЗАВИСИМОСТИ

Методические указания
для самостоятельной работы студентов

Составители И.А. Иконникова, Н.В. Лаходынова

Томск 2014

Введение в статистические зависимости /Сост. И.А. Иконникова, Н.В. Лаходынова. – Томск: Изд-во Том. гос. архит.-строит. ун-та, 2014. – 23 с.

Рецензент Р.И. Лазарева
Редактор О.В. Иванова

Методические указания к самостоятельной работе по дисциплине Б2.Б.1 – «Математика» при изучении темы «статистические зависимости» студентами второго курса очной и заочной форм обучения всех направлений всех специальностей и профилей подготовки специалистов и бакалавров.

Печатаются по решению методического семинара кафедры высшей математики, протокол № 5 от 07.03.2014 г.

срок действия с 01.09.2014 до 01.09.2019

Оригинал-макет подготовлен И.А. Иконниковой.

Подписано в печать 17.04.2014 г.
Формат 60×84. Бумага офсет. Гарнитура Таймс.
Уч.-изд. л. 2,37. Тираж 30 экз. Заказ № .

Изд-во ТГАСУ, 634003, г. Томск, пл. Соляная, 2.
Отпечатано с оригинал-макета в ООП ТГАСУ.
634003, г. Томск, ул. Партизанская, 15.

ОГЛАВЛЕНИЕ

1. ПРЕДИСЛОВИЕ

Предлагаемые методические указания предназначены для самостоятельной работы студентов заочного факультета в процессе выполнения контрольной работы по теме «Статистические зависимости». Математическое содержание данного раздела направлено на формирование у студента следующих общекультурных (ОК) и профессиональных компетенций (ПК):

(ОК-1): владение культурой мышления, способностью к обобщению, анализу, восприятию информации, постановке цели и выбору путей её достижения.

(ОК-6): стремление к саморазвитию, повышению своей квалификации и мастерства.

(ОК-9): способность к целенаправленному применению базовых знаний в области математических, естественных, гуманитарных и экономических наук в профессиональной деятельности.

(ОК-15): владение методами количественного анализа и моделирования, теоретического и экспериментального исследования.

(ПК-1): способность использовать законы и методы математики, естественных, гуманитарных и экономических наук при решении профессиональных задач.

(ПК-32): способность выбирать математические модели организационных систем, анализировать их адекватность, проводить адаптацию моделей к конкретным задачам.

В результате освоения материала студент должен:

Знать:	понятие связи случайных величин, варианты её математического моделирования.
Уметь:	использовать статистические методы в обработке экспериментальных данных.
Владеть:	методами теории вероятностей и математической статистики, методами корреляционного и регрессионного анализа.

2. ВВЕДЕНИЕ В ПРЕДМЕТ

Основная задача естествознания состоит в изучении зависимостей между переменными. Особая роль при этом отводится математическому моделированию, которое лежит в основе компьютерного моделирования и обработки информации. Математическая модель развивает наши представления о закономерностях изучаемого процесса, позволяет выявлять управляющие факторы, регулировать их, обеспечивая тем самым прогнозируемые результаты. Именно в этом её цель и прикладное значение.

Рассмотрим простой пример. Изучая процесс упругого удлинения тонкого стержня под действием приложенной силы, в 1660 году Р. Гуком получена линейная зависимость вида:

$$l = l_0 + k \cdot F.$$

Здесь l_0 - исходная, а l - конечная длина стержня, F – величина приложенной силы, k – коэффициент пропорциональности.

Полученная Гуком формула отвечала практической надобности в прогнозировании поведения реальных инженерных конструкций под нагрузкой.

Здесь важно отметить, что на практике используют не саму формулу $l = l_0 + k \cdot F$, а *математическую модель* на её основе. В чём разница?

С теоретической точки зрения формула $l = l_0 + k \cdot F$ имеет неограниченную область определения, так как является линейной функцией. Это значит, что она должна быть справедлива для любых значений аргумента F.

Напротив, в реальности при оценке изменения размеров металлических конструкций под нагрузкой формула Гука применима лишь в *ограниченном* диапазоне величин F. Действительно, за пределом упругости начинается фаза пластического течения материала с принципиально другими закономерностями и, следовательно, другими моделями.

Таким образом, в математической модели всегда присутствуют ограничения естественного характера, которые чётко обозначают область её практического применения (границы *адекватности* модели). Именно это обстоятельство отличает математическую модель от простой математической формулы.

Приведём ещё один аргумент в пользу изучения зависимостей между переменными с применением метода математического моделирования. В современных условиях решение этой, впрочем, и любой другой инженерной задачи, предполагает широкое использование компьютера, которое возможно только в рамках идеологии математического моделирования.

2.1. Основные виды связи между переменными

На содержательном уровне две или более переменных зависимы (связаны) между собой, если наблюдаемые значения этих переменных изменяются согласованным образом. Проще говоря, переменные X и Y связаны, если изменение одной из них вызывает изменение другой.

Интуитивно ясно, что чем теснее связь между переменными, тем больше информации содержит одна переменная относительно другой, тем точнее можно спрогнозировать неизвестное значение одной переменной по заданной величине другой.

Максимально полная и точная связь переменных свойственна детерминированным процессам, в этом случае текущее состояние процесса целиком и полностью определяется его предшествующим состоянием. Другими словами, связь, при которой с изменением одной переменной вторая изменяется *строго определённым образом*, называется *детерминированной* (*функциональной*). Например, площадь квадрата Y функционально связана с длиной его стороны X: каждому значению X соответствует строго определенное значение Y. Функциональную связь изучают в классических естественных науках (физике, математике, механике и т. д.). В реальности пример детерминированной связи найти довольно трудно.

Другой предельный вариант – полное отсутствие связи между независимыми переменными.

Промежуточный вариант *неполной, случайной связи* встречается на практике. Действительно, все реально наблюдаемые явления и их показатели являются случайными по своей природе. Это значит, что ход развития таких явлений и их результат существенно зависит от множества факторов, часть которых не поддаётся учёту и контролю. Как результат - почти все наблюдаемые показатели об-

наруживают характерный для случайных величин разброс значений даже в неизменных условиях испытания. Поэтому и взаимосвязь случайных величин всегда статистическая, то есть неполная.

Статистической называют зависимость, при которой изменение одной из величин влечет изменение распределения другой. В частном случае, когда при изменении одной из величин изменяется только *среднее* значение другой, а закон распределения сохраняется, статистическая зависимость называется *корреляционной*.

Далее будем рассматривать только корреляционную зависимость, которая является свободной, неполной и неточной и потому она проявляется при массовых испытаниях в виде *тенденции*.

Корреляция (*correlation* – согласование, связь, соотношение) как термин впервые введён Гальтоном в 1888 году.

Обобщим сказанное. Так как в реальных процессах все характерные для них переменные и связи между ними фиксируются с некоторыми ошибками, корректное изучение этих процессов следует проводить в рамках *статистического исследования зависимостей.*

2.2. Два этапа математического моделирования связи

Математическое моделирование статистической связи случайных переменных выполняется в два этапа.

Сначала нужно ответить на следующий вопрос: имеется ли вообще какая-либо связь между изучаемыми переменными? Методы и модели, привлекаемые на первом этапе, составляют содержание *корреляционного анализа*. По его результатам статистическое заключение имеет вид: “связь есть” или “связи нет”. Факт обнаружения связи сопровождается численной оценкой её тесноты (силы).

При положительном заключении корреляционного анализа (*связь есть и она достаточно тесная*) переходят ко второму этапу по определению формы, то есть формулы связи между переменными. Задача решается средствами *регрессионного анализа*.

Обозначим некоторые принципиальные моменты корреляционного и регрессионного анализа на простом примере двух случайных переменных Y и X. Предполагается, что для изучения их связи исследователь располагает эмпирическими (опытными) данными в виде выборочной совокупности пар (x_i, y_i), $i = 1, \ldots, n$.

2.3. Основы корреляционного анализа

При первичном анализе все пары наблюдений (x_i, y_i), $i = 1, \ldots, n$ изображают на координатной плоскости. В результате получают так называемое *корреляционное поле* (*диаграмму рассеяния*). Взаимное положение точек визуально характеризует форму взаимосвязи X и Y (рис. 1).

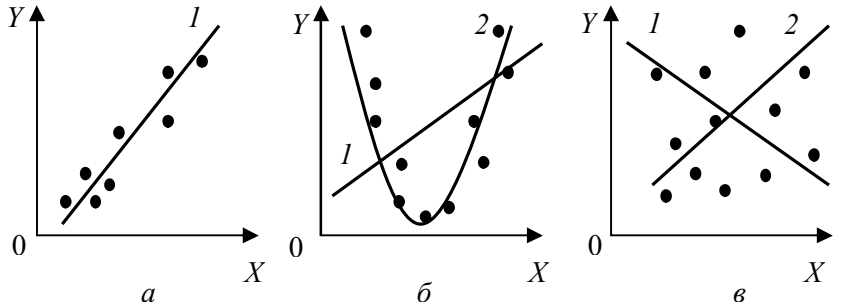

Рис. 1. Варианты взаимосвязи между X и Y

Так, прямая линия *1* на рис. 1, *а* достаточно хорошо представляет эмпирические данные, поэтому здесь между X и Y можно предположить *линейную* форму связи. Или, как говорят статистики, в изменении X и Y на рис. 1, *а* просматривается *линейная тенденция* (английский эквивалент – *тренд*) вида $Y = \beta_0 + \beta_1 X$.

На рис. 1, *б* форма связи X и Y близка к *квадратичной*, то есть тренд напоминает параболу. На рис. 1, *в* эмпирические данные имеют вид хаотического облака. В этом случае считается, что между X и Y отсутствует сколько-нибудь определённая взаимосвязь.

Итак, корреляционная связь случайных переменных X и Y проявляется следующим образом: каждому значению независимой переменной X соответствует не одно значение Y, а совокупность с некоторым средним по совокупности $M_x(Y)$. Именно это среднее значение $M_x(Y)$ меняется в зависимости от X, то есть

$$M_x(Y) = f(X).$$

Здесь $M_x(Y)$ – *условное математическое ожидание* величины Y,

соответствующее фиксированному значению x случайной величины X; функция $f(X)$, называемая *функцией регрессии*, она характеризует тенденцию в зависимости между X и Y.

Наибольшую практическую значимость имеет случай, когда связь между переменными X и Y близка к линейной (рис. 1, *а*), рассмотрим его подробнее.

Мерой тесноты линейной связи двух случайных переменных служит *коэффициент линейной корреляции Пирсона*, который по выборочным данным (x_i, y_i) $i = 1, \ldots, n$ оценивают по формуле:

$$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\cdot(y_i - \bar{y})}{S_x \cdot S_y} = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2}\cdot\sqrt{\overline{y^2} - \bar{y}^2}}, \qquad (1)$$

где $\bar{x}$, $\bar{y}$ – выборочные средние; S_x, S_y – выборочные средние квадратические отклонения (стандартные ошибки).

Корреляцию между двумя переменными следует понимать так: изменение (*вариация*) одной из них сопровождается изменением другой. При *линейной* корреляции формула (1) определяет коэффициент пропорциональности между взаимными вариациями X и Y.

Основные свойства выборочного коэффициента корреляции:

1. Коэффициент корреляции принимает численные значения на отрезке [–1,1], то есть $-1 \le r_{xy} \le 1$.

2. Если $r_{xy} > 0$, то между величинами X и Y есть положительная линейная корреляция. Это значит, что переменные изменяются в одном направлении: с ростом X величина Y *в среднем* возрастает, а снижение X приводит к снижению Y в среднем. Если же $r_{xy} < 0$, то корреляция отрицательная, то есть переменные X и Y изменяются в противоположных направлениях.

3. Если $r_{xy} = 0$, то между величинами X и Y нет *линейной* корреляционной связи (но, возможно, есть нелинейная корреляция).

4. Чем ближе r_{xy} по модулю к 1, тем теснее корреляция X и Y.

5. Если величина коэффициента корреляции из (1) определяет тесную связь X и Y, то уравнение тренда имеет линейный вид, то есть $Y = \beta_0 + \beta_1 X$.

Следует помнить, что коэффициент корреляции r_{xy} вычисляется по выборочным данным и может неправильно представлять

реальное положение дел. Например, ненулевое значение выборочного коэффициента корреляции r_{xy} может быть целиком обусловлено случайным колебанием выборки, на основе которой он вычислен.

В силу сказанного необходима проверка статистической значимости коэффициента корреляции. Для этого по выборке (x_1, y_1), (x_2, y_2), …, (x_n, y_n) объёма n вычисляют статистику Стьюдента:

$$t = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}, \qquad (2)$$

В зависимости от величины t из (2) выбираем один из следующих вариантов заключения:

- Если $1 < |\, t \,| \le 2$, то коэффициент корреляции статистически незначим, то есть между X и Y в реальности отсутствует линейная корреляционная связь.
- Если $2 <|\, t \,| \le 3$, то коэффициент корреляции статистически значим (доверительная вероятность лежит между значениями 0,95 и 0,99). Это значит, что между X и Y в действительности есть линейная корреляционная связь.
- Если $|\, t \,| > 3$, наличие линейной связи X и Y почти гарантировано.

Приведённый способ проверки значимости носит приближённый характер. Тем не менее, он довольно часто используется на практике, особенно в прикидочных оценках.

Если статистическая значимость коэффициента корреляции подтверждена, то оценку тесноты связи X и Y делают с использованием так называемой шкалы Чеддока (табл. 1).

Таблица 1

Величина $\lvert r_{xy} \rvert$	0,1 – 0,3	0,3 – 0,5	0,5 – 0,7	0,7 – 0,9	0,9 – 0,99
Характеристика силы связи	слабая	умеренная	заметная	высокая	весьма высокая

В специальной литературе часто встречается понятие “ложная корреляция”. Приведём классический пример А.А. Чупрова: стати-

стический анализ данных по множеству городов обнаружил положительную корреляцию между числом пожарных команд и величиной убытков от пожаров: чем больше пожарных, тем больше убытки. Такой результат противоречит здравому смыслу. Поэтому при более детальном изучении вопроса была найдена причина ложной корреляции. Она состоит в наличии третьей переменной (величина города), которая опосредует связь первых двух. Действительно, чем больше город, тем больше число пожарных команд, чем больше город, тем больше пожаров, то есть убытков. Именно эти пары переменных имеют между собой положительную связь.

Основная причина ложной корреляции состоит в том, что вы не знаете, чем она вызвана. Даже при высоком значении коэффициента корреляции связь между переменными должна быть обоснована по существу (то есть выявлен её причинный механизм).

Подытожим содержание данного раздела. Результаты корреляционного анализа отвечают на следующие вопросы:

1. Имеется ли связь между изучаемыми переменными?
2. Какова структура этой связи?
3. Как измеряется теснота этой связи?

Дальнейшее изучение взаимосвязи переменных, то есть получение формулы связи, считают целесообразным при $| r_{xy} | \geq 0{,}7$.

2.4. Содержание регрессионного анализа

Пусть между изучаемыми переменными X и Y надёжно установлена корреляционная связь, тогда методы регрессионного анализа позволяет найти формулу, описывающую эту связь.

В специальной литературе зависимую переменную Y называют *результирующей*, *объясняемой*, а независимую переменную X, которая определяет изменение Y, называют *объясняющей*, *факторной переменной*.

Обозначим некоторые принципиальные допущения (предпосылки) регрессионной модели.

Предполагается, что случайные по своей природе значения $Y(X)$ можно представить в виде суммы двух слагаемых, одно из которых представляет закономерную (то есть неслучайную)

часть в изменении Y от X, а другое – случайную. В результате имеем:

$$Y(X) = f(X) + \varepsilon(X).$$

Неслучайная компонента $f(X) = M_x(Y)$ – это функция регрессии, которая представляет тренд в зависимости Y от X. *Случайная* компонента $\varepsilon(X)$ отражает тот факт, что реальные значения зависимой переменной не всегда совпадают с её условным математическим ожиданием и могут быть различными при одном и том же значении объясняющей переменной.

Предполагается также, что природа случайной компоненты $\varepsilon(X)$ и характеристики её распределения никак не связаны со структурой функции регрессии $f(X)$.

Вспомним, что общий вид функции регрессии (уравнение тренда) устанавливают в ходе корреляционного анализа. Поэтому в изучаемом здесь случае *линейной* связи имеем $f(x) = \beta_0 + \beta_1 X$. Тогда с учётом случайной поправки $\varepsilon(X)$:

$$Y(X) = f(X) + \varepsilon(X) = \beta_0 + \beta_1 \cdot X + \varepsilon(X). \qquad (3)$$

Соотношение (3) называют *моделью парной линейной регрессии*.

Таким образом, регрессионная модель предлагает неизвестные истинные значения Y оценивать посредством функции регрессии, то есть аппроксимировать взаимосвязь X и Y трендом $f(X)$. В частности, линейная регрессия предполагает, что при массовых наблюдениях случайного показателя Y его значения изменяются в среднем по линейному закону. Результаты отдельных наблюдений незначительно отклоняются от линейного тренда на случайную величину $\varepsilon(X)$, которая даёт “ошибки модели” относительно тренда.

Параметризация модели парной линейной регрессии

Модель регрессии (как и любая другая теоретическая модель) содержит внутри себя множество параметров. Значения этих параметров “привязывают” модель к реальности каждой конкретной задачи, тогда как исходная теоретическая модель пригодна для описания целого класса однотипных задач. Расчётную базу для определения параметров модели составляют результаты наблюдений Y и

X, то есть выборочные данные.

Так, модель парной линейной регрессии (3) для каждого i-го наблюдения имеет вид:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \ . \qquad (4)$$

Параметры модели - постоянные β_0 и β_1 называются *теоретическими коэффициентами регрессии* и подлежат определению по выборочным данным.

Оценка теоретической модели (4) по выборочным данным, то есть *эмпирическая* (выборочная) модель имеет вид:

$$y_i = b_0 + b_1 x_i + e_i = \hat{y}_i + e_i \,, \qquad (5)$$

здесь b_0 и b_1 – выборочные оценки неизвестных теоретических коэффициентов β_0, β_1, тогда $\hat{y}_i = b_0 + b_1 x_i$ – выборочная оценка функции регрессии $f(x) = \beta_0 + \beta_1 X$, наконец, e_i – оценка случайной составляющей ε_i из (4).

На практике для вычисления b_0 и b_1 чаще всего используют метод наименьших квадратов (МНК), который проиллюстрирован на рис. 2.

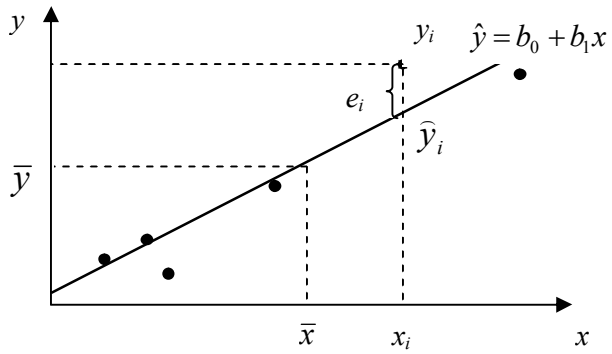

Рис. 2. Иллюстрация принципа МНК

Идея МНК заключается в следующем: из всего множества линий, которые можно провести через экспериментальные точки на корреляционном поле, линия регрессии выбирается так, чтобы *сумма квадратов погрешностей для всех точек* $(e_1^2 + e_2^2 + \ldots + e_n^2)$ *была наименьшей*.

Действительно, пусть e_i^2 - квадрат отклонения наблюдаемого значения y_i от его оценки по модели регрессии $\hat{y}_i$ в точке x_i (рис. 2), тогда, суммируя по всем точкам $i = 1,...,n$, получим:

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 \cdot x_i)^2 .$$

Величина Q оценивает суммарную погрешность аппроксимации выборочных данных (x_i, y_i), $i = 1, ..., n$ посредством линии регрессии $\hat{y} = b_0 + b_1 x$. Заметим, что функция $Q = Q(b_0, b_1)$, тогда условие её минимума:

$$\begin{cases} \dfrac{\partial Q}{\partial b_0} = -2\sum\limits_{i=1}^{n} (y_i - b_0 - b_1 \cdot x_i) = 0 \\ \dfrac{\partial Q}{\partial b_1} = -2\sum\limits_{i=1}^{n} (y_i - b_0 - b_1 \cdot x_i) \cdot x_i = 0 \end{cases}.$$

Отсюда

$$\begin{cases} nb_0 + b_1 \sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} y_i \\ b_0 \sum\limits_{i=1}^{n} x_i + b_1 \sum\limits_{i=1}^{n} x_i^2 = \sum\limits_{i=1}^{n} x_i \cdot y_i \end{cases}.$$

В результате применение МНК даёт для оценок b_0 и b_1 следующие формулы:

$$b_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{n\sum\limits_{i=1}^{n} x_i y_i - \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} y_i}{n\sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2}; \qquad (6)$$

$$b_0 = \bar{y} - b_1 \bar{x} = \frac{1}{n}\sum_{i=1}^{n} y_i - b_1 \frac{1}{n}\sum_{i=1}^{n} x_i . \qquad (7)$$

Угловой коэффициент регрессии b_1 показывает, на сколько единиц в среднем изменяется зависимая переменная Y при изменении независимой переменной X на единицу своего измерения. *Постоянная* b_0 дает среднее значение переменной Y при $X = 0$, то есть b_0 определяет точку пересечения линии регрессии с осью ординат.

Полученная в результате выборочная функция регрессии

$$\hat{y}_i = b_0 + b_1 x_i \qquad (8)$$

широко используется на практике для прогнозирования, так как даёт приближённую оценку $\hat{y}_i$ для неизвестного значения Y в любой интересующей нас точке x_i.

Отметим, что из двух параметров b_0 и b_1 в (8) более важен угловой коэффициент регрессии b_1, так как именно он связывает изучаемый признак Y с объясняющей переменной X. Если коэффициент b_1 получился близким к нулю, то в реальности Y скорей всего не зависит от X (или зависит, но не линейно, а, например, квадратично).

Вопрос значимости углового коэффициента регрессии изучен теоретически и в случае *парной линейной регрессии* доказано, что он значим (существенно отличен от нуля) при условии, что коэффициент корреляции оказался значимым.

Ещё одним важным параметром модели является *коэффициент детерминации*:

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}. \qquad (9)$$

Величина R^2 показывает, какую долю общей дисперсии результативного признака (знаменатель) объясняет уравнение регрессии (числитель). По способу определения R^2 изменяется в пределах от нуля до единицы: $0 \le R^2 \le 1$.

Для парной линейной регрессии (и только!) коэффициент детерминации равен квадрату коэффициента корреляции $R^2 = r_{xy}^2$, причём значимость коэффициента детерминации следует из значимости коэффициента корреляции.

Геометрический смысл определения (9) состоит в следующем: чем ближе R^2 к единице, тем теснее точки наблюдений примыкают к линии регрессии, тем лучше уравнение регрессии описывает исходные выборочные данные. Следовательно, величину коэффициента детерминации можно использовать для оценки качества модели регрессии: из двух предложенных моделей лучше та, у которой выше коэффициент детерминации. Действительно, в статистической практике зачастую предлагают несколько разных моделей для описания одних и тех же выборочных данных (*конкурс моделей*).

Побеждает та модель регрессии, у которой выше коэффициент детерминации.

Прогнозирование в регрессионной модели

Под прогнозированием понимают оценивание зависимой переменной y_p для заданного значения факторной переменной x_p, которого нет в исходных наблюдениях.

Точечный прогноз – это оценка $\hat{y}_p$ которая рассчитана по уравнению регрессии (8):

$$\hat{y}_p = b_0 + b_1 x_p \qquad (10)$$

Точность прогноза убывает по мере удаления x_p от $\bar{x}$ (это следует из формул для вычисления коэффициентов b_0 и b_1). С другой стороны, увеличение объёма выборки n способствует улучшению качества прогноза.

На практике уравнение регрессии считается пригодным для прогнозирования, если коэффициент детерминации $R^2 > 0{,}5$. В этом случае оно объясняет более половины дисперсии Y его зависимостью от X. Оставшаяся доля дисперсии Y, равная $(1 - R^2)$, объясняется либо случайными, либо не учтёнными в модели факторами.

Подытожим содержание данного раздела. Регрессионный анализ отвечает на следующие вопросы.

1. Какова общая структура математической модели, предназначенной для описания связи случайных переменных.
2. Как в этой модели провести конкретные оценки её параметров, исходя из имеющихся опытных (эмпирических) данных.
3. Каковы “деловые” (прогностические) качества построенной модели.

2.5. Заключение по линейной модели регрессии

Модель линейная регрессии, как и любая другая модель, нуждается в анализе её адекватности.

Во-первых, с этой целью используют коэффициент детерминации, так как он характеризует величину ошибки модели.

Во-вторых, адекватность модели регрессии, естественно, определяют статистические свойства её остатков e_i ($i = 1,\ldots,n$). Обо-

значим только общий план соответствующего исследования.

Регрессионная модель предлагает вместо неизвестных истинных значений $Y(x_i)$ использовать величину функции регрессии $\hat{y}_i$, вычисленную в той же точке x_i. Такая "подмена" тем успешней, чем лучше статистические свойства остатков $e_i = y_i - \hat{y}_i$ в точках x_i, $i = 1,\ldots,n$, (рис. 2). А именно, обусловленные конкретной моделью остатки должны подчиняются нормальному закону распределения с нулевым средним и постоянной дисперсией, а также быть статистически независимыми в разных наблюдениях. Подтверждение перечисленных свойств остатков служит доказательством того, что регрессионная модель построена оптимально, а остатки можно квалифицировать как "шум" или случайные помехи.

В заключение следует отметить, что линейная модель регрессии имеет исключительное теоретическое и практическое значение. А именно:

1. Линейная модель регрессии глубоко проработана в теоретическом плане. В силу этого она нашла реализацию во всех практически используемых статистических пакетах, нацеленных на компьютерную обработку экспериментальных данных (Excel, Mathcad, Statistica,…).
2. Модель проста как в реализации, так и в прикладном толковании.
3. Линейная модель может выступать в качестве начального приближения в процессе последовательного продвижения к более сложной и более адекватной модели.
4. Иногда интересующая нас с целью прогнозирования область может быть *локально* представлена линейной функцией (за пределами этой области данные наблюдений могут иметь нелинейный характер).
5. Довольно широкий класс нелинейных регрессий сводится к линейным путем тождественных математических преобразований. Это так называемые *линейные относительно параметров модели.* Выполнив, как правило, логарифмирование и/или замену переменных, можно получить линейную функцию регрессии.

3. ВОПРОСЫ ДЛЯ САМОПРОВЕРКИ

1. Понятие о математическом моделировании взаимосвязи переменных. Пример.
2. Основные виды связи между переменными.
3. Детерминированная связь переменных, пример.
4. Определение статистической связи переменных.
5. Основные этапы изучения статистической зависимости.
6. Понятие корреляционной связи.
7. Понятие ложной корреляции. Пример.
8. Понятие корреляционного поля.
9. Основные задачи корреляционного анализа.
10. Примеры вариантов корреляционной зависимости.
11. Выборочный коэффициент корреляции, основные свойства.
12. Проверка значимости выборочного коэффициента корреляции.
13. Что такое шкала Чеддока? Для чего она используется?
14. Варианты заключений по результатам корреляционного анализа.
15. Задача регрессионного анализа.
16. Общий вид модели парной линейной регрессии. Статистический смысл её компонент.
17. Идея метода наименьших квадратов (МНК).
18. Выборочная модель парной линейной регрессии.
19. Суть коэффициента детерминации, границы его изменения.
20. Прогнозирование по регрессионной модели.

4. ПОСТАНОВКА ЗАДАЧИ. ПРОЦЕДУРА РЕШЕНИЯ

Постановка задачи: из непрерывной генеральной совокупности двух случайных переменных X и Y произведена случайная выборка объёма n: (x_i, y_i) $i = 1, \ldots, n$. Изучить связь этих переменных с использованием модели парной линейной регрессии.

Процедура решения.

1. Построить корреляционное поле по выборочным данным.
2. Определить визуально пригодность линейной формы связи между X и Y.

3. Оценить тесноту линейной связи X и Y по величине коэффициента корреляции.
4. По выборочным данным оценить коэффициенты уравнения регрессии и записать его в явном виде. Определить величину коэффициента детерминации.
5. Рассчитать точечный прогноз модели.

5. ВАРИАНТЫ ИНДИВИДУАЛЬНЫХ ЗАДАНИЙ

Варианты индивидуальных заданий приведены в табл. 2. Выбор варианта производится по последней цифре номера зачетной книжки. Так, если номер зачетной книжки заканчивается цифрой восемь, то Ваш вариант № 8, и так далее. Если номер зачетной книжки заканчивается нулём, то Ваш вариант № 10.

Таблица 2

№ 1	*X*	74	86	115	125	150
	Y	0,5	0,7	1,2	1,5	2
№ 2	*X*	70	85	112	123	150
	Y	0,6	0,8	1,2	1,4	2
№ 3	*X*	74	88	115	120	145
	Y	0,5	0,8	1,2	1,5	1,9
№ 4	*X*	71	86	115	125	150
	Y	0,5	0,7	1,2	1,5	1,8
№ 5	*X*	75	86	115	125	150
	Y	0,5	0,7	1,2	1,5	2
№ 6	*X*	76	85	112	123	150
	Y	0,6	0,8	1,2	1,4	2
№ 7	*X*	77	88	115	120	145
	Y	0,5	0,8	1,2	1,5	1,9
№ 8	*X*	78	86	115	125	150
	Y	0,5	0,7	1,2	1,5	1,8
№ 9	*X*	69	86	115	125	150
	Y	0,5	0,7	1,2	1,5	2
№ 10	*X*	70	85	112	123	150
	Y	0,6	0,8	1,2	1,4	2

6. ПРИМЕР ВЫПОЛНЕНИЯ РАБОТЫ

Изучается связь между величиной расходов на питание (X) и размером прожиточного минимума (Y). Результаты наблюдений названных переменных приведены в табл. 3.

Найти прогнозное значение для размера прожиточного минимума y_p, если расходы на питание составляют $x_p = 20$ (усл. ед.).

Таблица 3

X	2	6	10	14	18
Y	9	10	12	19	20

1. Корреляционное поле представлено на рис. 3. Из постановки задачи ясно, что размер прожиточного минимума – результативный признак (Y), а величина расходов на питание – факторный признак (X).

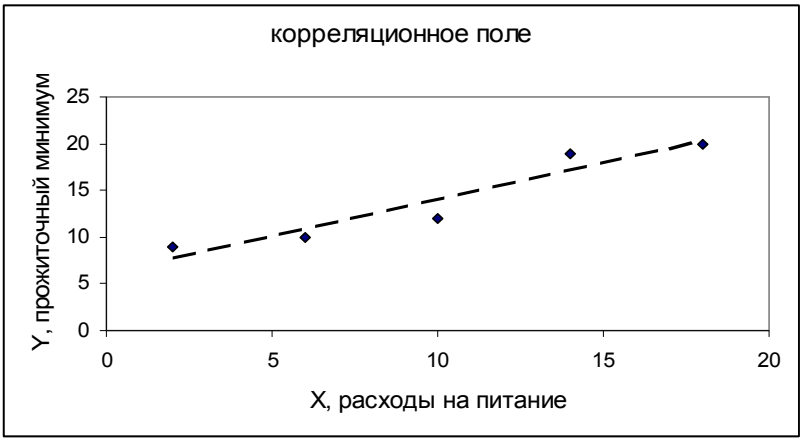

Рис. 3. Корреляционное поле для изучаемых переменных

2. По виду корреляционного поля (точки на рис. 3) заключаем, что линейная форма связи между Y и X вполне допустима, её возможный вариант изображён пунктиром на рис. 3.

3. Для дальнейших расчетов заготовим вспомогательную таблицу (табл. 4).

Первые два столбца содержат выборочные данные из постановки задачи. На их основе заготовлены значения, нужные для оценки коэффициента корреляции согласно формуле (1). В двух последних строках табл. 4 размещены суммарные и средние значения величин, необходимые для последующих расчётов.

Таблица 4

x	y	xy	x^2	y^2	$(x-\bar{x})^2$	$\hat{y} = b_0+b_1 x$	$e^2=(y-\hat{y})^2$
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
2	9	18	4	81	64	7,8	1,44
6	10	60	36	100	16	10,9	0,81
10	12	120	100	144	0	14,0	4,00
14	19	266	196	361	16	17,1	3,61
18	20	360	324	400	64	20,2	0,04
50	70	824	660	1086	160	70,0	9,90
10	14	164,8	132	217,2	64	14,0	1,98

Заметим, что при заполнении этой таблицы довольно часто ошибаются, считая, что $\overline{x^2} = \bar{x}^2$. Рекомендуем расписать несколько первых слагаемых при вычислении средних вручную, тогда легко убедиться в том, что $\overline{x^2} \neq \bar{x}^2$.

Используем данные первых пяти столбцов табл. 4, чтобы оценить тесноту линейной связи переменных Y и X по формуле (1):

$$r_{xy} = \frac{164{,}8 - 10 \cdot 14}{\sqrt{132 - 10^2} \cdot \sqrt{217{,}2 - 14^2}} = 0{,}952 .$$

Проверим статистическую значимость коэффициента корреляции.

Согласно (2), наблюдаемое значение t-статистики Стьюдента:

$$t = \frac{0{,}952 \cdot \sqrt{5-2}}{\sqrt{1 - 0{,}952^2}} \approx 5{,}39 .$$

Так как $|t| > 3$, коэффициент корреляции статистически значим.

Таким образом, между наблюдаемыми переменными есть положительная линейная связь: с увеличением расходов на питание

величина прожиточного минимума в среднем возрастает. Теснота линейной связи согласно шкале Чеддока (табл. 1) весьма высокая.

4. Для получения формулы, связывающей размер прожиточного минимума (Y) с величиной расходов на питание (X), построим выборочное уравнение регрессии.

Для этого найдём оценки коэффициентов уравнения регрессии по формулам (6) и (7). А именно, используя данные табл. 4, получим $b_0 = 6{,}25$, $b_1 = 0{,}775$.

Коэффициент b_1 показывает, что при увеличении расходов на питание на 1 усл. ед. прожиточный минимум увеличится в среднем на 0,775 усл. ед. Значение b_0 показывает, что при нулевом уровне расходов на питание прожиточный минимум составит в среднем 6,25 усл. ед. Знаки коэффициентов уравнения регрессии соответствуют логике изучаемого явления.

С учётом найденных коэффициентов выборочное уравнение регрессии имеет вид: $\hat{y} = 6{,}25 + 0{,}775 \cdot x$. Используя это уравнение, заполним столбцы 7 и 8 в табл. 4. Сравнение значений в столбце 2 (наблюдаемые, выборочные данные y_i, $i = 1, \ldots, n$) с соответствующими цифрами столбца 7 (значения $\hat{y}_i$, предсказанные моделью регрессии в тех же точках x_i, $i = 1, \ldots, n$) определяет остатки модели $e_i = y_i - \hat{y}_i$ в каждой точке наблюдений x_i, $i = 1,\ldots,n$, (рис. 2). Сумма квадратов этих величин (столбец 8) характеризует точность модели в целом.

Оценим качество модели, вычислив коэффициент детерминации по формуле (9). В нашем примере получим $R^2 = r_{xy}{}^2 \approx 0{,}91$, отсюда заключаем, что 91 % дисперсии результативного признака объясняется уравнением регрессии $\hat{y} = 6{,}25 + 0{,}775 \cdot x$, а доля необъяснённой дисперсии составляет лишь 9 %.

Близость величины коэффициента R^2 к единице свидетельствует о высоком качестве уравнения регрессии.

5. Найдем прогноз модели для уровня прожиточного минимума при величине расходов на питание $x_p = 20$ (усл. ед.).

Точечный прогноз получим в результате подстановки интересующего нас значения $x_p = 20$ в уравнение (10):

$$\hat{y}_p = 6{,}25 + 0{,}775 \cdot 20 = 21{,}75.$$

При расходах на питание в 20 усл. ед. наиболее вероятная величина прожиточного минимума будет 21,75 усл. ед.

Заключение: при исследовании связи между размером прожиточного минимума Y и величиной расходов на питание X по выборочным данным построено уравнение регрессии $\hat{y} = 6{,}25 + 0{,}775 \cdot x$, которое имеет хорошие статистические свойства и, следовательно, может быть рекомендовано для практического применения.

7. БИБЛИОГРАФИЧЕСКИЙ СПИСОК

Основная литература

1. Гмурман, В.Е. Теория вероятностей и математическая статистика / В.Е Гмурман. – М. : Высшая школа, 2011. – 480 с.
2. Сидняев, Н.И. Теория планирования эксперимента и анализа статистических данных / Н.И. Сидняев. – М.: Юрайт, 2011. – 399 с.
3. Воскобойников, Ю.Е. Регрессионный анализ данных в пакете Mathcad / Ю.Е. Воскобойников. – М.: Лань, 2011. – 223 с.
4. Мамонтов, Г.Я. Анализ данных в MS EXCEL и OO CALC / Г.Я. Мамонтов, И.А. Иконникова. – Томск: Офсетная лаборатория ТГАСУ, 2010. – 59 с.
5. Иконникова, И.А. Эконометрика / И.А. Иконникова, Н.А. Вихорь. – Томск: Офсетная лаборатория ТГАСУ, 2012. – 87 с.

Дополнительная литература

6. Айвазян, С.А. Прикладная статистика. Основы эконометрики / С.А. Айвазян, В.С. Мхитарян. - М.: Юнити, 1998. – 1022 с.
7. Айвазян, С.А. Прикладная статистика: Исследование зависимостей / С.А. Айвазян, И.С. Енюков, Л.Д. Мешалкин. – М.: Финансы и статистика, 1985.- 487 с.
8. Тюрин, Ю.Н. Анализ данных на компьютере / Ю.Н. Тюрин, А.А. Макаров. - М.: ИД «ФОРУМ», 2008. – 368 с.
9. Протасов, К.В. Статистический анализ экспериментальных данных / К.В. Протасов. – М.: Мир, 2005. – 142 с.
10. Просветов, Г.И. Анализ данных с помощью Excel: задачи и решения / Г.И. Просветов. – М.: Альфа-Пресс, 2009. – 157 с.