Министерство образования Республики Беларусь
БЕЛОРУССКИЙ НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Кафедра «Высшая математика № 3»

В.В. Веременюк
В.В. Кожушко
Е.А. Крушевский

# УЧЕБНО-МЕТОДИЧЕСКОЕ ПОСОБИЕ

к лабораторной работе № 2
«Установление зависимости между двумя случайными величинами по результатам их выборок»

М и н с к 2 0 0 4

УДК 519.216 (075.8)
ББК 22.1 я 7
В 31

Рецензенты:
А.Д. Корзников, Т.Н. Гурина

**Веременюк В.В**

В 31 Учебно-метод пособие к лабораторной работе № 2 «Установление зависимости между двумя случайными величинами по результатам их выборки» / В.В. Веременюк, В.В. Кожушко, Е.А. Крушевский. – Мн.: БНТУ, 2004. – 45 с.

ISBN 985-479-092-4.

Учебно-методическое пособие содержит теоретический материал, необходимый для выполнения лабораторной работы «Установление зависимости между двумя случайными величинами по результатам их выборок». Четвертый раздел издания содержит дополнительный материал по теме «Множественная линейная регрессия».

ISBN 985-479-092-4

## Лабораторная работа № 2

# УСТАНОВЛЕНИЕ ЗАВИСИМОСТИ МЕЖДУ ДВУМЯ СЛУЧАЙНЫМИ ВЕЛИЧИНАМИ ПО РЕЗУЛЬТАТАМ ИХ ВЫБОРОК

## Цель работы

На практике часто необходимо исследовать, как изменение одной величины ($X$) влияет на другую величину ($Y$). Например, как количество цемента ($X$) влияет на прочность бетона ($Y$). Такое влияние иногда может описываться простой функциональной связью $Y = F(X)$ между переменными. Однако для многих изучаемых процессов это скорее исключение, чем правило. Тем не менее исследователь все-таки отмечает некую существенную связь между переменными. Эта так называемая корреляционная связь и будет предметом нашего изучения в данной лабораторной работе. На основе статистического анализа полученных экспериментальных данных, которые будут представлены, как правило, в виде таблицы чисел ($x_1$, $y_1$), ($x_2$, $y_2$), ..., ($x_n$, $y_n$), студент должен научиться следующему:

– устанавливать наличие или отсутствие связи (корреляционной) между изучаемыми величинами $X$ и $Y$;

– предсказывать тип зависимости между переменными $X$ и $Y$, т.е. выдвигать модель исследуемой связи, (как правило, это будет полином не слишком высокой степени);

–оценивать параметры предложенной модели, например коэффициенты соответствия полинома;

– проверять адекватность построенной модели реальному процессу, т.е. овладеть процедурой проверки гипотез о значимости упомянутых коэффициентов.

## Содержание работы

1. Установление наличия корреляционной зависимости между случайными величинами $X$ и $Y$:

– определение выборочного коэффициента корреляции по результатам выборки (эксперимента);

– установление значимости выборочного коэффициента корреляции;

– выводы о зависимости или независимости случайных величин $X$ и $Y$.

2. Выбор регрессионной модели и ее статистический анализ:

– выбор уравнения регрессии;

– оценка коэффициентов регрессионного уравнения по методу наименьших квадратов;

– проверка точности оценки регрессии и установление адекватности выбранной модели изучаемому процессу.

## Порядок проведения работы

Изучить теоретический материал.

По данным выборки найти выборочный коэффициент корреляции $r_{xy}$.

Установить значимость отличия от нуля $r_{xy}$.

Если сделан вывод о том, что между $X$ и $Y$ существует корреляционная связь, выбрать уравнение линии (модель) регрессии $Y$ на $X$.

Оценить параметры выбранной модели по результатам выборки.

Проверить точность оценки регрессии, т.е. найти доверительные интервалы для параметров регрессии.

Провести расчеты на ПЭВМ.

Сделать основные выводы.

Составить отчет по работе.

### Требования к отчету

Отчет по работе должен состоять из следующих разделов:
Постановка задачи.
Анализ решения задачи.
Результаты счета на ПЭВМ.
Основные выводы и рекомендации.

## 1. УСТАНОВЛЕНИЕ НАЛИЧИЯ ЗАВИСИМОСТИ МЕЖДУ ДВУМЯ СЛУЧАЙНЫМИ ВЕЛИЧИНАМИ ПО РЕЗУЛЬТАТАМ ИХ ВЫБОРОК.

В предыдущей лабораторной работе мы изучали одну случайную величину $X$, ее вероятностные и статистические характеристики, полученные на основе имеющейся выборки значений. Большой интерес для широкого класса научных и инженерных задач представляет обнаружение взаимных связей между двумя и более случайными величинами. Например, существует ли связь между курением и ожидаемой продолжительностью жизни или между умственными способностями и успеваемостью. Очевидно, что привычной нам строгой функциональной зависимости, когда каждому значению $X$ по определенному правилу соответствует значение $Y$ т.е., $Y = f(X)$, здесь мы установить не можем. Слишком много случайных факторов влияют как на величину $X$, так и на величину $Y$. Тем не менее зачастую невооруженным глазом видно, что какая-то зависимость между изучаемыми величинами $X$ и $Y$ существует. Например, вес и рост человека, естественно, тесно связаны между собой, но они не определяют друг друга однозначно. Точно так же прочность бетона, очевидно, зависит от количества цемента, но эта зависимость явно не функциональная.

Рассмотрим задачу, как сделать надежный вывод о наличии или отсутствии зависимости между двумя случайными величинами на основе экспериментальных данных. Слово «надежный»

означает (как и в лабораторной работе № 1), что вероятность истинности сделанного вывода должна быть близка к единице.

Для исследования такой зависимости во второй половине XIX века английский ученый Фрэнсис Гальтон (двоюродный брат Чарльза Дарвина) и его ученики (например, Карл Пирсон) ввели такие важные понятия, как корреляция и регрессия, которые стали основными понятиями в теории вероятностей и математической статистике, а также в связанных с ними научных дисциплинах. При этом саму зависимость между случайными величинами назвали *корреляционной*.

## 1.1. Коэффициент корреляции

В инженерных приложениях задача о наличии зависимости между двумя случайными величинами обычно сводятся к установлению связи между некоторыми предполагаемыми возбуждениями $x_1, x_2, .., x_n$ и наблюдаемым откликом $y$ изучаемой системы(рис. 1.1).

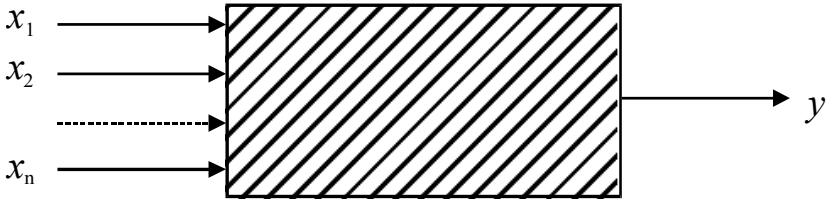


Рис. 1.1

Рассмотрим вначале взаимосвязь между одним фактором x и откликом $y$, т.е. взаимосвязь между двумя случайными величинами $X$ и $Y$. В теории вероятностей для установления степени *корреляционной зависимости* между двумя случайными величинами вводится числовая характеристика, которая называется коэффициентом корреляции или просто *корреляцией.* Для двух случайных величин $X$ и $Y$ коэффициент корреляции определяется следующим образом: где

$$\rho_{xy} = \rho(X,Y) = \frac{M[(X - M_x)(Y - M_y)]}{\sigma_x \sigma_y},$$

$M_x, \sigma_x$ и $M_y, \sigma_y$ математическое ожидание и среднеквадратическое отклонение $X$ и $Y$ соответственно.

Величину, которая стоит в числителе правой части этой формулы, называют *ковариацией* и часто обозначают $C(X,Y)$:

$$C(X,Y) = M[(X - M_x)(Y - M_y)].$$

Легко увидеть, что если $X = Y$, то

$$C(X,X) = \sigma_x^2, \quad C(Y,Y) = \sigma_y^2.$$

При этом выполняется неравенство

$$|C(X,Y)| \le \sigma_x \sigma_y.$$

Это означает, что коэффициент корреляции заключен между – 1 и 1:

$$-1 \le \rho(X,Y) \le 1.$$

$$M[(X - M_x)(Y - M_y)] = M[XY - XM_x - YM_y + M_xM_y] =$$

$$= M(XY) - M_xM_y - M_yM_x + M_xM_y = M(XY) - M_xM_y,$$

формулу для коэффициента корреляции (1.1) легко преобразовать в более наглядную и удобную в вычислениях: Учитывая, что

$$\rho(X,Y)=\frac{M(XY)-M_x M_y}{\sigma_x \sigma_y}.$$

Если случайные величины $X$ и $Y$ – независимы, то

$$M(X\cdot Y)=M(X)\cdot M(Y),$$

коэффициент корреляции $\rho(X,Y)=0$. Обратное утверждение, вообще говоря, неверно, т.е. *некоррелированные случайные величины не обязательно независимы*.

Определенный выше коэффициент корреляции принадлежит к числу наиболее трудных для понимания числовых характеристик случайных величин. Если такие характеристики, как частота, вероятность, математическое ожидание, дисперсия, стандартное отклонение, осознаются достаточно легко, то термин "коэффициент корреляции" оказывается сложным для понимания. Это объясняется в первую очередь сложностью математического выражения для этого коэффициента и отсутствием соответствующего понятия в повседневной жизни. Поэтому напомним основные свойства коэффициента корреляции:

1. Коэффициент корреляции $\rho(X,Y)$ симметричен относительно $X$ и $Y$ и может изменяться в пределах от – 1 до 1.

2. Равенство $|\rho(X,Y)|=1$ указывает на наличие точной линейной связи вида $y=\rho\frac{\sigma_y}{\sigma_x}(x-M_x)+M_y$ между рассматриваемыми величинами $X$ и $Y$, возможные значения которых в этом случае расположены на одной прямой.

3. При значениях $\rho(X,Y)$, по модулю близких к единице,, точки с координатами $X$ и $Y$ с большой вероятностью располагаются в окрестностях некоторой прямой, непараллельной ни одной из осей координат. По мере уменьшения $|\rho(X,Y)|$ эта вероятность падает, и при $\rho(X,Y)$, близком к нулю, такая прямая вообще не может быть обнаружена. Однако это не

исключает возможности нелинейной связи между рассматриваемыми величинами.

Таким образом, коэффициент корреляции является характеристикой степени и направления линейной связи между двумя величинами. В заключение отметим, что:

– при $|\rho(X,Y)|=1$ величины $X$ и $Y$ являются полностью коррелированными; при $|\rho(X,Y)|$, близком к единице, – сильно коррелированными; при $\rho(X,Y)$, близком к нулю, – слабо коррелированными; при $\rho(X,Y)=0$ – некоррелированными;

– при $\rho(X,Y)>0$ корреляционная связь между величинами $X$ и $Y$ положительная; при $\rho(X,Y)<0$ – отрицательная. При этом в первом случае возрастанию $X$ с большой вероятностью соответствует возрастание $Y$, а во втором – убывание.

## 1.2. Представление исходных данных.

Приступая к анализу экспериментальных данных, мы не имеем сведений о наличии коррелированной связи между случайными величинами $X$ и $Y$. Поэтому первая задача, которую необходимо решить, – это сделать обоснованный вывод о наличии или об отсутствии этой зависимости. Как и в случае исследования одной случайной величины, материалом, на базе которого решаем эту задачу, является выборка значений случайной величины $X$ и $Y$.

Пусть в результате n экспериментов (проведенных в одних и тех же условиях) зафиксированы n пар значений $(x_1, y_1)$, $(x_2, y_2)$,..., $(x_n, y_n)$, где $x_i$ - значение, которое приняла случайная величина $X$ в $i$-м эксперименте, а $y_i$ – значение, которое приняла случайная величина $Y$ в этом же эксперименте. Этот набор данных называется выборкой объема n пары случайных величин.

В случае большого объема выборки ($n$ – достаточно велико) и частого повторения в выборке некоторых значений $x_i$ и $y_i$ ее удобно представлять в виде *корреляционной таблицы*:

| $Y$ \ $X$ | $x_1$ | $x_2$ | … | $x_k$ | |
|---|---|---|---|---|---|
| $y_1$ | $m_{11}$ | $m_{12}$ | … | $m_{1k}$ | $m_{Y_1}$ |
| $y_2$ | $m_{21}$ | $m_{22}$ | … | $m_{2k}$ | $m_{Y_2}$ |
| … | … | … | | … | … |
| $y_l$ | $m_{l1}$ | $m_{l2}$ | … | $m_{lk}$ | $m_{Y_l}$ |
| | $m_{X_1}$ | $m_{X_2}$ | … | $m_{X_k}$ | |

Здесь $m_{ij}$ – *частота* появления пары значений $(x_i, y_j)$ в выборке; $x_1 < x_2 < ... < x_k$ – различные значения, которые приняла случайная величина $X$, а $m_{X_1},...,m_{X_k}$ – частоты появления этих значений в выборке $(\sum_{i=1}^{k} m_{X_i} = n)$. Аналогично, $y_1 < y_2 < ... < y_k$ –значения, которые приняла случайная величина $Y$, $m_{Y_1},...,m_{Y_l}$ – частоты появления этих значений в выборке $(\sum_{i=1}^{l} m_{Y_i} = n)$.

Если при большом объеме выборки также велико число различных значений $x_i$ и $y_i$, то вначале таблицу имеет смысл представить в виде интервальной корреляционной таблицы (аналог группированного статистического ряда, см. лабораторную работу № 1):

| $Y$ \ $X$ | $[x_0, x_1)$ | $[x_1, x_2)$ | … | $[x_{k-1}, x_k)$ |
|---|---|---|---|---|
| $[y_0, y_1)$ | $m_{11}$ | $m_{12}$ | … | $m_{1k}$ |
| $[y_1, y_2)$ | $m_{21}$ | $m_{22}$ | … | $m_{2k}$ |
| … | … | … | | … |
| $[y_{l-1}, y_l)$ | $m_{l1}$ | $m_{l2}$ | … | $m_{lk}$ |

Здесь частота $m_{ij}$ – количество пар значений $(x_\zeta, y_\theta)$ выборки, для которых значение $x_\zeta$ попадает в интервал $[x_{i-1}, x_i)$, а значение $y_\theta$ – в интервал $[y_{j-1}, y_j)$. Количество интервалов, их длина, а также начальные значения $x_0, y_0, x_k, y_l$ определяются так же, как и для группированного статистического ряда (см. лабораторную работу № 1, с. 22 – 23).

Следует отметить, что часто информация для статистического исследования сразу поступает в виде интервальной корреляционной таблицы.

Для дальнейших вычислений информацию, представленную в виде интервальной корреляционной таблицы, надо преобразовать в обычную корреляционную таблицу, взяв в качестве значений вариант середины соответствующих интервалов:

| $Y$ \ $X$ | $x_1^*$ | $x_2^*$ | … | $x_k^*$ |
|---|---|---|---|---|
| $y_1^*$ | $m_{11}$ | $m_{12}$ | … | $m_{1k}$ |
| $y_2^*$ | $m_{21}$ | $m_{22}$ | … | $m_{2k}$ |
| … | … | … | | … |
| $y_l^*$ | $m_{l1}$ | $m_{l2}$ | … | $m_{lk}$ |

Здесь $x_i^* = \frac{x_i + x_{i-1}}{2}$, $y_j^* = \frac{y_j + y_{j-1}}{2}$.

Разберем сказанное на примере. Пусть получена выборка случайных величин $X$ и $Y$ ($X$ – рост студента в см; $Y$ – его масса в кг).

Данные о росте и массе студентов ($X$ – рост в см; $Y$ – масса в кг):

| $X$ | 178 | 188 | 178 | 165 | 175 | 185 | 183 | 175 | 183 | 193 | 188 | 183 | 173 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $Y$ | 63 | 95 | 67 | 66 | 83 | 75 | 70 | 77 | 79 | 70 | 84 | 84 | 75 |

| $X$ | 178 | 180 | 173 | 185 | 165 | 185 | 188 | 163 | 183 | 183 | 170 | 185 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $Y$ | 100 | 84 | 82 | 77 | 61 | 79 | 82 | 68 | 77 | 75 | 66 | 77 |

Объем выборки $n = 25$. Найдем количество интервалов по формуле $k \geq [\log_2 n] + 1$ (см. лабораторную работу № 1, с. 33): $k \geq [\log_2 25] + 1 = 4 + 1 = 5$.

Итак, можно взять $k = 5$. Так как $x_{min} = 163$ и $x_{max} = 193$, то длина интервала для случайной величины $X$ будет равна

$$h_x = \frac{193 - 163}{5} = 6,$$

$$x_0 = 163,$$

$$x_5 = 193.$$

Аналогично $y_{min} = 61,\ y_{max} = 100,\ h_y = \frac{100 - 61}{5} = 7{,}8$. Округлим $h_y$ до 8 и возьмем $y_0 = 60,\ y_5 = 100$. Построим интервальную корреляционную таблицу:

| $Y$ \ $X$ | [163;169) | [169;175) | [175;181) | [181,187) | [187,193) |
|---|---|---|---|---|---|
| [60;68) | 2 | | 3 | | |
| [68;76) | 1 | 1 | | 3 | 1 |
| [76;84) | | 1 | 2 | 5 | 1 |
| [84;92) | | | 1 | 1 | 1 |
| [92;100) | | | 1 | | 1 |

Преобразуем ее в обычную корреляционную таблицу:

| $Y$ \ $X$ | 166 | 172 | 178 | 184 | 190 | $m_{X_i}$ |
|---|---|---|---|---|---|---|
| 64 | 2 | | 3 | | | 5 |
| 72 | 1 | 1 | | 3 | 1 | 6 |
| 80 | | 1 | 2 | 5 | 1 | 9 |
| 88 | | | 1 | 1 | 1 | 3 |
| 96 | | | 1 | | 1 | 2 |
| $m_{Y_j}$ | 3 | 2 | 7 | 9 | 4 | 25 |

Пустые клетки означают, что соответствующая им частота равна 0.

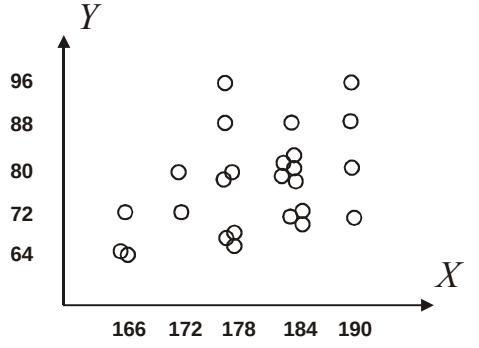


Рис. 1.2

Для большей наглядности исходные данные представляют в виде так называемого *корреляционного поля.* (рис. 1.2). Для этого в системе координат *XOY* строят точки $(x_i, y_j)$, причем их количество соответствует частоте из корреляционной таблицы. При частоте, большей единицы, изображается облако близко расположенных друг к другу точек с соответствующими координатами.

### 1.3. Оценка коэффициента корреляции по результатам эксперимента

Пусть имеется выборка значений пары случайных величин *X* и *Y*, которая представлена в виде набора пар ($x_1$, $y_1$), ($x_2$, $y_2$),..., ($x_n$, $y_n$) или в виде корреляционной таблицы.

По аналогии с оценками математического ожидания и дисперсии будем утверждать, что достаточно хорошей оценкой коэффициента корреляции $\rho(X,Y)$ будет *выборочный коэффициент корреляции,* который обозначим $r(X,Y)$ и будем вычислять по формуле

$$r(X,Y) = \hat{\rho}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{[\sum_{i=1}^{n}(x_i - \overline{X})^2 \sum_{i=1}^{n}(y_i - \overline{Y})^2]^{1/2}}.$$

Заметим, что $r_{XY}$ – случайная величина, которая принимает конкретные значения при некоторой реализации совокупности (*X, Y*): ($x_1$, $y_1$), ...,($x_n$, $y_n$).

Вычисления выборочного значения $r_{XY}$ коэффициента корреляции удобно использовать формулу

$$r_{XY} = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\hat{\sigma}_x \hat{\sigma}_y},$$

где $\overline{XY} = \frac{1}{n}\sum_{i=1}^{n} x_i\, y_i$, $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} x_i$,

$$\hat{\sigma}_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{X}^2}, \quad \hat{\sigma}_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{Y}^2}.$$

или, если исходные данные представлены в виде корреляционной таблицы,

$$\overline{XY} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{l} m_{ij} x_i^* y_j^*,$$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{k} m_{X_i} x_i^*,$$

$$\overline{Y} = \frac{1}{n}\sum_{j=1}^{l} m_{Y_j} y_j^*$$

$$\hat{\sigma}_x = \sqrt{\frac{1}{n}\sum_{i=1}^{k} m_{X_i} (x_i^*)^2 - \overline{X}^2}$$

$$\hat{\sigma}_y = \sqrt{\frac{1}{n}\sum_{j=1}^{l} m_{Y_j} (y_j^*)^2 - \overline{Y}^2}.$$

Как и $\rho(X,Y)$, выборочный коэффициент корреляции лежит между – 1 и 1 и принимает одно из граничных значений только при наличии идеальной линейной связи между наблюдениями. Нелинейная связь и/или разброс данных, вызванный ошибками измерения или же неполной коррелированностью случайных величин, приводит к уменьшению абсолютного

значения $r(X,Y)$. Характерные виды корреляционных полей для различных значений коэффициента корреляции приведены на рис. 1.3.

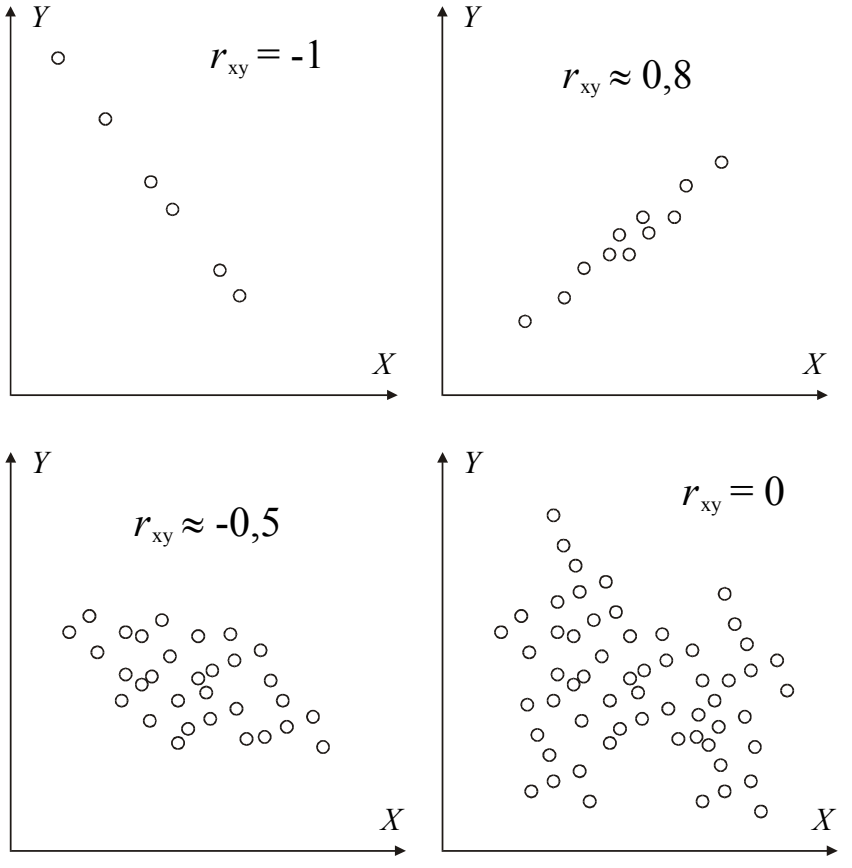


Рис. 1.3

## 1.4. Проверка значимости выборочного коэффициента корреляции

По методике, изложенной в предыдущем разделе, используя выборочные значения совокупности случайных величин

*(X, Y)*: ($x_1$, $y_1$), ..., *($x_n$, $y_n$)* найдем *выборочный коэффициент корреляции* $r(X,Y)$. Так как выборка отобрана случайно, то можно утверждать, что $r(X,Y)$ лишь приближенно представляет коэффициент корреляции $\rho(X,Y)$ случайных величин *X* и *Y*. Но в конечном итоге нас интересует именно коэффициент $\rho(X,Y)$, поэтому хотелось бы знать, насколько можно доверять полученной оценке. Другими словами, ставится *задача о значимости найденного выборочного коэффициента корреляции.* Эту задачу можно сформулировать следующим образом: необходимо при заданном уровне значимости $\alpha$ проверить нулевую гипотезу $H_0$: $\rho(X,Y) = \rho_0$ против альтернативной гипотезы $H_1$: $\rho(X,Y) \neq \rho_0$, если известно, что выборочный коэффициент корреляции $r(X,Y)$ принял значение, неравное нулю.

На практике чаще приходится встречаться с проверкой этой гипотезы при $\rho_0 = 0$. Эта задача связана с решением вопроса о зависимости или независимости случайных величин X и Y. Действительно, если гипотеза $H_0$: $\rho(X,Y) = 0$ отвергается, то это значит, что выборочный коэффициент корреляции *значимо отличается от нуля* и можно сделать вывод, что между *X* и *Y* существует статистически значимая корреляционная связь. Если же нулевая гипотеза будет принята, то выборочный коэффициент корреляции незначим, а *X* и *Y* некоррелированы, т.е. не связаны какой-либо зависимостью.

Итак, решается задача в следующей постановке: проверить нулевую гипотезу $H_0$: $\rho(X,Y) = 0$ против альтернативной гипотезы $H_1$: $\rho(X,Y) \neq 0$.

Для проверки этой гипотезы, как и ранее, используется t-распределение (распределение Стьюдента).

Рассмотрим следующую случайную величину:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

(сравним со случайной величиной $t = \frac{\overline{X} - a}{\frac{S}{\sqrt{n-1}}}$ при $a = 0$ и $S = \sqrt{1-r^2}$ ).

Оказывается, во-первых, что эта случайная величина имеет распределение Стьюдента с $k = n - 2$ степенями свободы. Во-вторых, легко понять, почему эта случайная величина хорошо отражает изменения r: если $r \to 0$, то $t \to 0$, а если $|r| \to 1$, то $t \to \infty$. Таким образом, можно сказать, что гипотеза $H_0$ будет верна с вероятностью $p = 1 - \alpha$, если $P(|t| < t_\gamma) = 1 - \alpha$. Графически наблюдаемое значение должно попасть в область принятия гипотезы (рис. 1.4).

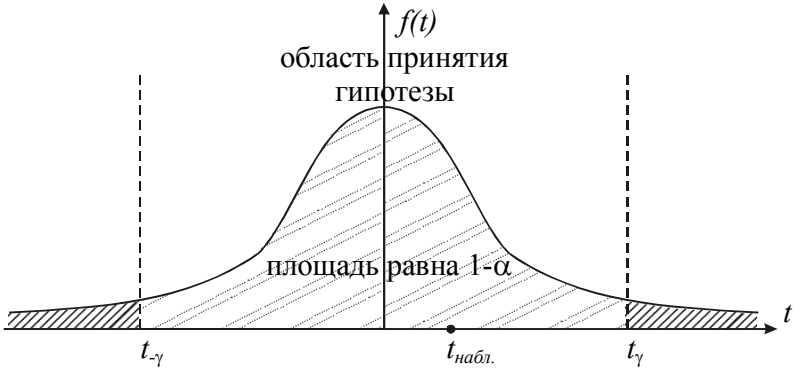


Рис. 1.4

Как и ранее, после несложных преобразований можно показать, что $t_\gamma$ находится по таблицам распределения Стьюдента из равенства:

$$P(t < t_\gamma) = f(t_\gamma) = 1 - \frac{\alpha}{2},$$

где *f(t)* – распределение Стьюдента с $n - 2$ степенями свободы. Таким образом, порядок шагов таков: вычисляем по результатам выборки значение

$$t_{\text{набл.}} = \frac{r_{XY}}{\sqrt{1 - r_{XY}^2}} \sqrt{n-2}$$

и сравниваем его с квантилем $t_\gamma$, найденным по таблицам распределения Стьюдента по вероятности $\gamma = 1 - \frac{\alpha}{2}$ и числу степеней свободы $k = n - 2$. Если $|t_{\text{набл.}}| > t_\gamma$, то гипотеза $H_0$ отвергается, т.е. считается, что $\rho \neq 0$, а $X$ и $Y$ коррелированы. Если $|t.| < t_\gamma$, то нет оснований отвергать гипотезу $H_0$, т.е. гипотезу о некоррелированности случайных величин.

### 1.5. Пример вычисления коэффициента корреляции

Составим таблицу о росте и массе для $n = 25$ выбранных наугад студентов. Есть ли основание считать, что рост и масса студентов коррелированы при уровне значимости $\alpha = 0{,}05$?

Данные о росте и массе студентов ($X$ – рост в см $Y$ – масса в кг):

| $X$ | 178 | 188 | 178 | 165 | 175 | 185 | 183 | 175 | 183 | 193 | 188 | 183 | 173 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $Y$ | 63 | 95 | 67 | 66 | 83 | 75 | 70 | 77 | 79 | 70 | 84 | 84 | 75 |

| $X$ | 178 | 180 | 173 | 185 | 165 | 185 | 188 | 163 | 183 | 183 | 170 | 185 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $Y$ | 100 | 84 | 82 | 77 | 61 | 79 | 82 | 68 | 77 | 75 | 66 | 77 |

Вычисляем следующие величины:

$$\sum_{i=1}^{25} x_i y_i = 344493,\ \sum_{i=1}^{25} x_i^2 = 806105,\ \sum_{i=1}^{25} y_i^2 = 148918,$$

$$\overline{X} = \frac{\sum_{i=1}^{25} x_i}{n} = \frac{4485}{25} = 179{,}4,$$

$$\overline{Y} = \frac{\sum_{i=1}^{25} y_i}{n} = \frac{1916}{25} = 76{,}64,\ r_{XY} =$$

$$= \frac{344493 - 25 \cdot 179{,}4 \cdot 76{,}64}{\sqrt{(806105 - 25 \cdot 179{,}4^2)(148918 - 25 \cdot 76{,}64^2)}} = 0{,}43.$$

По таблицам $t$ – распределения при уровне значимости α = 0,05 и числу степеней свободы $k = n - 2 = 23$ находим квантиль $t_{\gamma, n-2} = 1{,}714$.

Вычисляем: $t_{\text{набл.}} = \frac{r}{\sqrt{1 - r^2}} \cdot \sqrt{n-2} = \frac{0{,}43}{\sqrt{1 - 0{,}43^2}} \cdot \sqrt{23} = 2{,}29$.

Так как $t_{\text{набл.}} > t_{\gamma}$, то гипотеза $H_0$ должна быть отвергнута с уровнем значимости 5%. Следовательно, есть основания считать, что между ростом и массой студентов существует значимая корреляция.

## 2. ОСНОВЫ ЛИНЕЙНОГО РЕГРЕССИОННОГО АНАЛИЗА.

Корреляционный анализ позволяет установить степень взаимосвязи двух случайных величин. В случае, когда коэффициент корреляции $\rho(X, Y) \neq 0$, т.е. между $X$ и $Y$ существует значимая корреляционная связь, желательно иметь модель

этой связи, которая дала бы возможность предсказывать значения одной случайной величины $Y$ по конкретным значениям другой $X$ (или наоборот – $X$ по $Y$).

Например, корреляционный анализ данных, приведенный в предыдущем разделе, установил значимую линейную связь между ростом и массой студентов. Логичен следующий шаг: конкретизировать эту связь так, чтобы по данному росту можно было предсказать массу студента. Методы решения подобных задач изучает регрессионный анализ.

Коэффициент корреляции описывает зависимость между двумя случайными величинами одним числом, а регрессия выражает эту зависимость в виде функционального соотношения, поэтому дает более полную информацию. Так, регрессией является средний вес тела человека как функция от его роста.

Слово "регрессия" в статистику ввел, как мы уже упоминали Френсис Гальтон, один из создателей математической статистики. Сопоставляя рост детей и их родителей, он обнаружил, что соответствие между ростом отцов и детей слабо выражено, оно оказалось меньшим, чем он ожидал. Однако Гальтон объяснил это явление наследственностью не только от родителей, но и от более отдаленных предков: по его предположениям, т.е по его математической модели, рост определяется наполовину родителями, на четверть дедом и бабкой, на одну восьмую прадедом и прабабкой и т.д. Мы не знаем, прав ли Гальтон, но он обратил внимание на движение назад по генеалогическому дереву и назвал это явление регрессией, заимствовав понятие движения назад, противоположное прогрессу – движению вперед. Впоследствии слово "регрессия" заняло в статистике заметное место, хотя, как это часто бывает в любом языке, в том числе и в языке науки, в него теперь вкладывают другой смысл – оно означает функциональную статистическую связь между случайными величинами.

## 2.1. Функции и линии регрессии

Что же понимают под словом "регрессия" или "функциональной статистической связью" между случайными величинами в теории вероятностей? Пусть при каждом фиксированном значении случайной величины $X = x$, случайная величина $Y$ имеет определенное распределение вероятностей с условным математическим ожиданием, которое является функцией $x$: $M(Y / X = x) = f(x,\theta)$, где через θ обозначим совокупность $(\theta_1,\theta_2,...,\theta_k)$ параметров, определяющих функцию $f$. Так, в случае линейной зависимости

$$\theta = (a,b),\ f(x,a,b) = ax + b\ .$$

*Зависимость f (x, θ), где x – независимая переменная, называется регрессией или функцией регрессии случайной величины X на случайную величину Y.*

График функции $f(x, \theta)$ называется линией регрессии $Y$ на $X$.

Точность, с которой линия регрессии передает изменение $Y$ в среднем при изменении $X$, измеряется дисперсией величины $Y$, вычисляемой для каждого значения $x$:

$$D(Y / X = x) = \sigma^2(x)\ .$$

Линии регрессии обладают следующим замечательным свойством: среди всех действительных функций $f(x)$ минимум математического ожидания $M[(Y - f(x))^2]$ достигается для функции регрессии:

$$f(x,\theta) = M(Y / X = x)\ .$$

То есть регрессия $Y$ на $X$ дает наилучшее (в указанном смысле) представление величины $Y$.

Пусть $f(x, \theta)$ – функция регрессии $Y$ на $X$. Тогда зависимость между исследуемыми признаками (случайными величинами) $Y$ и $X$ запишем в виде $y = f(x,\theta) + \varepsilon$, где $\varepsilon$ – случайная величина, которую иногда называют ошибкой эксперимента, причем $M(\varepsilon) = 0,\ D(\varepsilon) = \sigma^2(x)$.

В дальнейшем будем рассматривать модели с постоянной дисперсией: $D(\varepsilon) = \sigma^2 = \text{const}$.

К сожалению, на практике законы распределения $X$ и $Y$ неизвестны, и функция регрессии определяется приближенно по экспериментальным данным методом регрессионного анализа. Цель регрессионного анализа состоит в определении общего вида уравнения регрессии, построении оценок неизвестных параметров, входящих в уравнение регрессии, и в проверке статистических гипотез, связанных с регрессией.

## 2.2. Модель линейного регрессионного анализ

Статистический подход к задаче построения (точнее, восстановления) функциональной зависимости $Y$ от $X$ основывается на предположении, что нам известны некоторые исходные (экспериментальные) данные $(x_i, y_i)$, $i = 1, 2, ..., n$, где $y_i$ – значения $Y$ при заданном значении $x_i$ - независимой переменной $X$, влияющей на значения $Y$. Пару значений $(x_i, y_i)$ часто называют результатом одного измерения, а $n$ – числом измерений; $Y$ – откликом, а $X$ – фактором, влияющим на отклик.

Предположим, что наблюдаемое в опыте значение отклика $Y$ можно мысленно разделить на две части: одна закономерно зависит от $X$, т.е. является функцией $X$, другая часть – случайна по отношению к $X$. Обозначим, как и ранее, первую часть через $f(x, \theta)$, вторую – через $\varepsilon$ и представим отклик в виде: $Y = f(x,\theta) + \varepsilon$, где $\varepsilon$ – случайная величина (ошибка эксперимента или измерения).

Применяя это соотношение к имеющимся у нас исходным данным, получим: $y_i = f(x_i,\theta) + \varepsilon_i, i = 1,2,...,n.$

Разделение $y_i$ на закономерную и случайную составляющую можно провести только мысленно. Реально ни $f(x_i)$, ни $\varepsilon_i$ в отдельности нам не известны, из опыта мы узнаем только их сумму – $y_i$. В связи с этим необходимо сделать определенные уточнения относительно величин $\varepsilon_i$. В классической модели регрессионного анализа предполагается, что:

– все опыты были проведены независимо друг от друга в том смысле, что случайности, вызвавшие отклонение отклика от закономерности в одном опыте, не оказывали влияния на подобные отклонения в других опытах;

–статистическая природа случайных составляющих оставалась неизменной во всех опытах.

Из этих предположений, очевидно, вытекает, что величины $\varepsilon_i \quad i = 1, 2, ..., n$ характеризуют ошибки, независимые при различных измерениях и одинаково распределенные с нулевым средним и постоянной дисперсией.

### 2.3. Общая схема решения регрессионных задач

Самый простой случай регрессионных задач – это исследование связи между одной независимой переменной *X* и одной зависимой переменной (откликом) *Y*. Эта задача носит название простой регрессии. Исходными данными являются два набора наблюдений: $x_1, x_2, ..., x_n$ – значения *X* и $y_1, y_2, ..., y_n$ – соответствующие значения *Y*. Перечислим основные этапы при решении задач простой регрессии.

Первым шагом решения задачи являются предположения о возможном виде функциональной связи между *X* и *Y*. Примерами возможных видов функциональных связей могут являться зависимости: $y = a + bx$,

$$y = a + bx + cx^2,\ y = e^{a+bx},\ y = 1/(a + bx)$$

и т.д., где *a, b, c* – неизвестные параметры, которые надо определить по исходным данным.

Для подбора вида зависимости между *X* и *Y* полезно построить и изучить расположение точек с координатами $(x_1, y_1)$, $(x_2, y_2)$,..., $(x_n, y_n)$.

Иногда примерный вид зависимости бывает известен из теоретических знаний или предыдущих исследований, аналогичным данным.

Важно выбрать функцию $f(x,\theta_1,...,\theta_k)$ так, чтобы она не просто хорошо описывала закономерную часть отклика, но и имела "физический" смысл, т.е. открывала объективную закономерность. Впрочем, полезны бывают и чисто эмпирические, "подгоночные" формулы, поскольку они позволяют в сжатой форме приближенно выразить зависимость *Y* от *X*.

## 3. ОЦЕНКА ПАРАМЕТРОВ МОДЕЛИ

После того как общий вид регрессионной функции выбран – $f(x,\theta_1,\theta_2,...,\theta_k)$, нужно подобрать оценки $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k$ соответствующих параметров таким образом, чтобы величины $r_i = y_i - f(x_i,\hat{\theta}_1,\hat{\theta}_2,...,\hat{\theta}_k)$ в совокупности были близки к нулю.

Значение $r_i$ называются остатками или отклонениями наблюдаемых величин $y_i$ от предсказанных

$$\hat{y}_i = f(x_i,\hat{\theta}_1,\hat{\theta}_2,...,\hat{\theta}_k).$$

Меру близости к нулю совокупности этих величин можно выбирать по-разному (например, максимум моделей, сумма модулей и др.). Но наиболее простые формулы расчета получаются, если в качестве этой меры выбрать сумму квадратов остатков (отклонений).

Если обозначим

$$S(\theta_1,\theta_2,...,\theta_k)=\sum_{i=1}^{n}[\,y_i-f(x_i,\theta_1,\theta_2,...,\theta_k)]^2\,,$$

то оценки параметров $\hat{\theta}_1,\hat{\theta}_2,...,\hat{\theta}_k$ выбираются из условия минимума суммы квадратов остатков $\min\limits_{\theta} S(\theta_1,\theta_2,...,\theta_k)$ или

$$\frac{\partial S}{\partial\theta_1}=0,\ \frac{\partial S}{\partial\theta_2}=0,\ ...,\ \frac{\partial S}{\partial\theta_k}=0\,.$$

Этот метод носит название метода наименьших квадратов.

*Методом наименьших квадратов* называется способ подбора параметров регрессионной модели на основе минимизации суммы квадратов остатков.

Сам по себе метод наименьших квадратов не связан с какими-либо предположениями о распределении случайных ошибок $\varepsilon_1,\varepsilon_2,...,\varepsilon_n$. Он может применяться и тогда, когда мы не считаем эти ошибки случайными (например, в задачах сглаживания экспериментальных данных).

Определив оценки параметров $\hat{\theta}_1,\hat{\theta}_2,...,\hat{\theta}$, мы тем самым определяем случайную величину $\hat{Y}=f(X,\hat{\theta}_1,\hat{\theta}_2,...,\hat{\theta}_k)$, которая называется среднеквадратической регрессией *Y* на *X*, а соответствующее уравнение $y=f(x,\hat{\theta}_1,\hat{\theta}_2,...,\hat{\theta}_k)$ называется выборочным уравнением функциональной регрессии *Y* на *X*.

*Замечания* 1. Уравнение регрессии *Y* на *X* позволяет предсказывать значения *Y* согласно выборочному уравнению регрессии $y=f(x,\hat{\theta}_1,\hat{\theta}_2,...,\hat{\theta}_k)$.

Если в исходных данных какому-либо значению $x_i$ соответствует несколько значений *Y*: $y_1, y_2, ..., y_m$, то

$$f(x_i,\hat{\theta}_1,\hat{\theta}_2,...,\hat{\theta}_k)\approx M(\hat{Y}/X=x_i)=\overline{y}_{x=x_i}\,.$$

Поэтому выборочное уравнение регрессии $Y$ на $X$ записывают иногда в виде

$$\overline{y}_x = f(x, \hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k).$$

2. Если в вышеприведенных распределениях случайные величины $X$ и $Y$ поменять местами, то аналогичные рассуждения приведут к уравнению регрессии $X$ на $Y$:

$$x = g(y, \hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k).$$

3. Широко распространенная в практических задачах ситуация, когда функция регрессии $f(x, \theta_1, \theta_2, ..., \theta_k)$ линейно зависит от параметров $\theta_1, \theta_2, ..., \theta_k$, носит название линейного регрессионного анализа. Например,

$$f(x, a_0, a_1, ..., a_k) = a_0 + a_1 x + ... + a_k x^k.$$

4. Ситуация, в которой экспериментатор может выбрать значения факторов $x_i$ по своему желанию и таким образом планировать будущие эксперименты, называется активным экспериментом. В этом случае значения факторов $x_i$ обычно рассматриваются как неслучайные. Более того, сообразуясь с целями эксперимента, экспериментатор может выбрать его план наилучшим образом (планирование эксперимента).

В отличие от этой ситуации в пассивном эксперименте значения фактора складываются вне воли экспериментатора, под действием других обстоятельств. Поэтому значения $x_i$ приходится толковать как случайные величины, что накладывает особые черты на интерпретацию результатов. Сама же математическая обработка совокупности $(x_i, y_i)$, $i = 1, 2, ..., n$, от этого не меняется.

## 3.1. Анализ адекватности модели

После подбора регрессионной модели и нахождения ее параметров желательно выяснить, насколько хорошо выбранная модель описывает имеющиеся данные. К сожалению, единого правила для этого нет. На практике первое впечатление о правильности подобранной модели могут дать изучение некоторых числовых характеристик, например доверительных интервалов для оценок параметров модели. Однако эти показатели скорее позволяют отвергнуть совсем неудачную модель, чем подтвердить правильность выбора функциональной зависимости.

Более обоснованные решения можно принять, сравнив имеющиеся значения $y_i$ со значениями $\hat{y}_i$, полученными с помощью подобранной функции регрессии $\hat{y}_i = f(x_i, \hat{\theta})$, т.е. провести анализ остатков $r_i = y_i - \hat{y}_i, i = 1, 2, ..., n$.

Исследование остатков полезно начинать с изучения их графика, который может показать наличие зависимости, не учтенной в модели. Скажем, при подборе простой линейной зависимости между *x* и *y* график остатков может показать необходимость перехода к нелинейной модели (квадратичной, полиномиальной, экспоненциальной). Для проверки нормальности распределения остатков чаще используется график плотности нормального распределения или критерии хи-квадрат Колмогорова и др.

## 3.2. Простейшая линейная регрессия

Проиллюстрируем изложенные выше идеи обработки регрессионного эксперимента на примере простой линейной регрессии. Допустим, что на первом этапе на основе анализа данных эксперимента, с учетом физических, экономических и других аспектов, а также прошлого опыта мы выбрали в качестве модельного уравнения регрессии прямую линию

$$f(x,a,b) = a + bx\,, \qquad (3.1)$$

т. е. в качестве модели регрессии между исследуемыми величинами $X$ и $Y$ берется линейная зависимость $Y = a + bX + \varepsilon$, Тогда для данного $x_i$ соответствующее значение $y_i$ определяется равенством $y_i = a + bx_i + \varepsilon_i, i = 1,2,...,n$, где $x_1$, $x_2$, ..., $x_n$ – заданные числа (значения фактора); $y_1$, $y_2$, ..., $y_n$ – наблюденные значения отклика; $\varepsilon_1$, $\varepsilon_2$,…, $\varepsilon_n$ – ошибки эксперимента, т.е. значения независимых, одинаково распределенных случайных величин.

Рассмотрим классическую модель регрессионного анализа, так называемую Гауссовскую модель, в которой предполагается, что величины $\varepsilon_i$ распределены по нормальному закону $N(0;\sigma^2)$ с некоторой неизвестной дисперсией $\sigma^2$.

Отметим, что предложенный вид зависимости и сделанные предположения насчет распределения остатков $\varepsilon_i$ – это модель, которой мы задаемся, но это не значит, что она верна. Начав с предположения, что эта модель установлена, на последующих стадиях анализа проверим адекватность модели реальному процессу и данным, полученным в результате наблюдения. Если факты будут против выбранной модели, то мы должны ее отклонить и разработать (выдвинуть) уже с учетом имеющейся информации другую модель и провести ее проверку.

На втором этапе анализа по результатам $n$ экспериментальных данных (наблюдений) пары величин $(X,Y)$: $(x_1, y_1)$, $(x_2, y_2)$,..., $(x_n,y_n)$ оценим параметры выдвинутой модели. В случае прямолинейной регрессии это $a$ и $b$.

Для более наглядной интерпретации результатов и упрощения расчетных формул преобразуем модельное уравнение регрессии, введя новый, неизвестный параметр $A = a + b\overline{X}$. Формула (3.1) примет вид

$$f(x, A, b) = A + b(x - \overline{X}).$$

Тогда предполагаемая связь между $x_i$ и $y_i$ запишется следующим оброзом:

$$y_i = A + b(x_i - \overline{X}) + \varepsilon_i, i = 1, 2, ..., n,$$

где $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$, т.е. фактически, мы ищем уравнения прямой регрессии в форме $y - y_0 = k(x - x_0)$.

Таким образом, *задача* – найти оценки параметров $A$ и $b$: $\alpha = \hat{A}, \ \beta = \hat{b}$, наилучшие в смысле метода наименьших квадратов, поскольку "истинная" линия регрессии, как известно, минимизирует математическое ожидание квадрата отклонения

$$[Y - f(X, \theta)]^2.$$

Теперь эмпирическое (выборочное) уравнение прямой регрессии $Y$ на $X$ запишем в виде $\hat{y} = \alpha + \beta(x - \overline{X})$.

Левая часть равенства, которую мы обозначили $\hat{y}$, является оценкой (приближенным значением) математического ожидания $M(Y)$ при заданном значении $x$. Поэтому выборочные уравнения прямой регрессии $Y$ на $X$ записывается иногда в виде $\hat{y}_x = \alpha + \beta(x - \overline{X})$.

Полученное уравнение можно использовать как предсказывающее: подстановка в него значения x позволяет "предсказать" среднее значение $Y$ для этого $x$. Если данные связаны идеальной линейной зависимостью ($|r_{XY}| = 1$), то предсказанные значения $Y$ будут в точности равняться наблюдаемому значению $y$ при данном $x$. Однако на практике обычно отсутствует идеальная линейная зависимость между данными, и

внешние случайные воздействия приводят к разбросу данных. Тем не менее если все же предположить существование линейной связи и наличие неограниченной выборки, то можно подобрать такие значения α и β, которые помогут предсказать ожидаемое значение $Y$ для любого значения $x$. Следовательно, значение $\hat{y}$ не обязательно совпадает с наблюдаемым значением $Y$, соответствующим данному $x$, однако оно будет равно среднему значению всех таких наблюдаемых значений.

Таким образом, на втором этапе исследования перед нами стоит задача: используя метод наименьших квадратов получить расчетные формулы для оценки параметров $A$ и $b$ прямолинейной регрессии.

*Замечание*. При изложении регрессионного анализа, как вы успели заметить, встает проблема в обозначениях: $x$ или $X$, у или $Y$ и др. Как и ранее, когда речь идет о вычислительных процедурах, обработке данных, линиях регрессии, изучении функциональной зависимости мы, обозначаем переменные через $x, y, \hat{y}$ . Например,

$$\hat{y} = \alpha + \beta(x - \overline{x}).$$

Если же нужно провести статистический анализ этого соотношения, то запишем так $\hat{Y} = \hat{A} + \hat{b}(X - \overline{X})$, т.е. как зависимость между случайными величинами. При изложении материала мы не оговариваем каждый раз, какие обозначения применяются, но надеемся, что из контекста это совершенно ясно.

## 3.3. Определение параметров прямолинейной регрессии методом наименьших квадратов

Суть метода наименьших квадратов состоит в том, что оценки $\hat{A}$ и $\hat{b}$ параметров $A$ и $b$ в предлагаемой линии регрессии $f(x, A, b) = A + b(x - \overline{x})$ подбирают таким образом, чтобы минимизировать сумму квадратов отклонений:

$$S(A,b) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [y_i - A - b(x_i - \bar{x})]^2 .$$

Данная функция принимает минимальное значение в точке, где обе частные производные обращаются в ноль:

$$\frac{\partial S}{\partial A} = 0, \quad \frac{\partial S}{\partial b} = 0 .$$

После дифференцирования получим:

$$\begin{cases} \dfrac{\partial S}{\partial A} = -2\sum_{i=1}^{n} [y_i - A - b(x_i - \bar{x})] = 0, \\ \dfrac{\partial S}{\partial b} = -2\sum_{i=1}^{n} (x_i - \bar{x})[y_i - A - b(x_i - \bar{x})] = 0. \end{cases}$$

После несложных преобразований имеем систему двух линейных уравнений:

$$\begin{cases} An + b\sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i-1}^{n} y_i , \\ A\sum_{i=1}^{n} (x_i - \bar{x}) + b\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i - \bar{x}) y_i . \end{cases}$$

Решив ее, получим искомые оценки $\hat{A}$ и $\hat{b}$:

$$\hat{A} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i ; \quad \hat{b} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} .$$

Данные оценки обладают следующими важными свойствами:

1. $M(\hat{A}) = A, M(\hat{b}) = b$.

2. $D(\hat{A}) = \sigma^2 / n, D(\hat{b}) = \sigma^2 \left( \sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^{-1}$.

3. $C(A,b) = 0$.

4. Случайные величины $\hat{A}$ и $\hat{b}$ распределены по нормальному закону.

5. $\hat{A}$ и $\hat{b}$ независимы как случайные величины.

Доказательствам утверждений 1 – 3 могут быть получены прямыми вычислениями, причем эти свойства не обязательно предполагают нормальный характер ошибок. Свойство 4 верно только в рассматриваемой нами Гауссовской модели. Свойство 5 есть естественное следствие нормальности ошибок и свойства 3 (если случайные величины, имеющие нулевой коэффициент ковариации, равны нулю, то они независимы). Независимость оценок $\hat{A}$ и $\hat{b}$ заметно упрощает дальнейший анализ.

*Замечания* 1. Полученные формулы для оценок $\hat{A}$ и $\hat{b}$ легко преобразовать к виду, более удобному для вычислений и анализа:

$$\hat{A} = \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i, \quad \hat{b} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x^2}, \text{ где } \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i, \quad \overline{xy} = \frac{1}{n} \sum_{i=1}^{n} x_i y_i, \ \sigma_x^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\bar{x}\right)^2 = \overline{x^2} - \left(\bar{x}\right)^2.$$

2. Уравнение регрессии $Y$ на $X$ записывается в виде

$$\hat{y} = \bar{y} + \hat{b}(x - \bar{x}).$$

Если $X$ и $Y$ – случайные величины, то, поменяв в наших выкладках местами $X$ и $Y$, получим прямую регрессию $X$ на $Y$:

$$\hat{x} = \bar{x} + \hat{b}_1 (y - \bar{y}).$$

где $$\hat{b}_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_y^2}, \quad \sigma_y^2 = \frac{\sum_{i=1}^{n} y_i^2}{n} - \left(\bar{y}\right)^2 = \overline{y^2} - \left(\bar{y}\right)^2.$$

Как видно, обе прямые регрессии проходят через точку $(\bar{x}, \bar{y})$. Угловые коэффициенты наклона прямых связаны с выборочным коэффициентом корреляции соотношением $r_{XY} = \sqrt{\hat{b} \cdot \hat{b}_1}$. Если учесть, что $\frac{\hat{b}}{\hat{b}_1} = \frac{\sigma_x^2}{\sigma_y^2}$, то можно получить выражение для оценок $\hat{b}$ и $\hat{b}_1$ через выборочный коэффициент корреляции и выборочные дисперсии

$$\hat{b} = \frac{\sigma_y}{\sigma_x} r_{XY}, \quad \hat{b}_1 = \frac{\sigma_x}{\sigma_y} r_{XY}.$$

Часто эти оценки называют выборочными коэффициентами регрессии $Y$ на $X$ и $X$ на $Y$ соответственно и обозначают $\rho_{Y/X}$ и $\rho_{X/Y}$, т.е.

$$\rho_{Y/X} = \hat{b} = \frac{\sigma_y}{\sigma_x} r_{XY}, \quad \rho_{X/Y} = \hat{b}_1 = \frac{\sigma_x}{\sigma_y} r_{XY}.$$

3. Легко увидеть, что прямые регрессии $Y$ на $X$ и $X$ на $Y$ совпадают только в том случае, если $|r_{XY}| = 1$, т.е. $X$ и $Y$ связаны линейной

зависимостью. Действительно, оба уравнения в этом случае преобразовываются к виду $\frac{y-\bar{y}}{\sigma_y}=\frac{x-\bar{x}}{\sigma_x}$.

## 3.4. Доверительные интервалы для параметров линейной регрессии

Свойства 1 – 4 оценок $\hat{A}$ и $\hat{b}$ параметров линейной регрессии показывают, что случайные величины $\hat{A}$ и $\hat{b}$ распределены по нормальному закону, причем

$$\hat{A}\cong N\left(A,\frac{\sigma^2}{n}\right),\quad \hat{b}\cong N\left(b,\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right),$$

где

$$D(\hat{A})=\sigma_{\hat{A}}^2=\frac{\sigma^2}{n},\quad D(\hat{b})=\sigma_{\hat{b}}^2=\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2},$$

$$M(\hat{A})=A,\quad M(\hat{b})=b.$$

Это дает возможность применить к построению доверительных интервалов ту же методику, что и для оценок неизвестного математического ожидания.

Если дисперсия ошибок эксперимента $\sigma^2$ – известна (что бывает крайне редко, и этот случай представляет собой больше теоретический интерес), то рассматриваем случайные величины

$$u_{\hat{A}} = \frac{\hat{A} - A}{\sigma_{\hat{A}}}, \quad u_{\hat{b}} = \frac{\hat{b} - b}{\sigma_{\hat{b}}},$$

которые имеют нормальное распределение $N(0;1)$. Для данного уровня значимости $\alpha$ получаем: $P(|u| < u_{1-\alpha/2}) = 1 - \alpha$. Из последнего соотношения находим $u_{1-\alpha/2}$ – квантиль нормального распределения, тогда

$$\frac{|\hat{A} - A|}{\sigma_{\hat{A}}} < u_{1-\alpha/2}, \quad \frac{|\hat{b} - b|}{\sigma_{\hat{b}}} < u_{1-\alpha/2}.$$

При этом доверительные интервалы будут следующими:

$$\hat{A} - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2} < A < \hat{A} + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2},$$

$$\hat{b} - \frac{\sigma}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}} u_{1-\alpha/2} < b < \hat{b} + \frac{\sigma}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}} u_{1-\alpha/2}$$

Пусть теперь $\sigma^2$ – неизвестная величина, что чаще бывает на практике. В таком случае необходимо воспользоваться оценкой $\hat{\sigma}^2$. Ключ к оцениванию $\sigma^2$ дает остаточная сумма квадратов:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left[ y_i - \hat{A} - \hat{b}(x_i - \bar{x}) \right]^2$$

Можно доказать, что в рассматриваемой нами Гауссовской модели эта сумма не зависит от $\hat{A}$ и $\hat{b}$ и имеет распределение

$\sigma^2 \chi^2_{n-2}$, где $\chi^2_{n-2}$ – распределение хи-квадрат с $n-2$ степенями свободы. Благодаря этому свойству для $\sigma^2$ можно построить несмещенную оценку $S_{yx}$:

$$S^2_{yx} = \frac{1}{n-2}\sum_{i=1}^{n}\left[y_i - \hat{A} - \hat{b}(x_i - \bar{x})\right]^2 .$$

Поскольку $S^2_{yx}$ не зависит от $\hat{A}$ и $\hat{b}$, то статистики $t_{\hat{A}} = \sqrt{n}\,\dfrac{\hat{A} - A}{\mathrm{S_{yx}}}$ и $t_{\hat{b}} = \dfrac{\hat{b} - b}{\mathrm{S_{yx}}}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ имеют распределение Стьюдента с $n-2$ степенями свободы. Тогда для данного уровня значимости $\alpha$ и по числу степеней свободы $n-2$ по таблицам квантилей распределения Стьюдента находим квантиль порядка $1-\frac{\alpha}{2}$, т.е. $t_{1-\alpha/2}$.

Доверительные интервалы для $\hat{A}$ и $\hat{b}$ запишутся в той же форме, что и при известном $\sigma^2$:

$$\hat{A} - \frac{S_{yx}}{\sqrt{n}}t_{1-\alpha/2} < A < \hat{A} + \frac{S_{yx}}{\sqrt{n}}t_{1-\alpha/2},$$

$$\hat{b} - \frac{S_{yx}}{\sqrt{n}\cdot\sigma_x}t_{1-\alpha/2} < b < \hat{b} + \frac{S_{yx}}{\sqrt{n}\cdot\sigma_x}t_{1-\alpha/2}$$

*Замечание.* Полученные выражения для доверительных интервалов можно записать в другой форме. Путем несложных преобразований и, с учетом того, что $\hat{b} = \dfrac{\sigma_y}{\sigma_x}r_{XY}$, $\hat{A} = \bar{y}$, остаточная сумма квадратов запишется в виде

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left[ y_i - \hat{A} - \hat{b}(x_i - \bar{x}) \right]^2 = n \cdot \sigma_y^2 (1 - r_{XY}^2).$$

Тогда $S_{yx} = \sigma_y \sqrt{\frac{n}{n-2}(1 - r_{XY}^2)}$. Кроме того, так как

$$\hat{\sigma}_x = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}, \quad \hat{\sigma}_y = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2},$$

то доверительные интервалы имеют вид

$$\hat{A} - \sigma_y \sqrt{\frac{1 - r_{XY}^2}{n-2}} t_{1-\alpha/2} < A < \hat{A} + \sigma_y \sqrt{\frac{1 - r_{XY}^2}{n-2}} t_{1-\alpha/2}, \quad (3.2)$$

$$\hat{b} - \frac{\sigma_y}{\sigma_x} \sqrt{\frac{1 - r_{XY}^2}{n-2}} t_{1-\alpha/2} < b < \hat{b} + \frac{\sigma_y}{\sigma_x} \sqrt{\frac{1 - r_{XY}^2}{n-2}} t_{1-\alpha/2}. \quad (3.3)$$

Последние формулы наиболее удобны для вычислений.

**Пример**. Определить по данным, приведенным в п. 1.2, прямую регрессии, задающую линейный прогноз средней массы студента по его росту. Найти 95%-й доверительный интервал для параметров прямой регрессии.

***Решение.*** С учетом вычислений, проделанных в п. 1.2 имеем, учитывая, что $n = 25$, $\alpha = 0{,}05$:

$$\hat{A} = \bar{y} = 76{,}64, \quad \hat{b} = \frac{344493 - 25 \cdot 179{,}4 \cdot 76{,}64}{806105 - 25 \cdot 179{,}4^2} = 0{,}51.$$

Следовательно, прямая регрессии, оценивающая среднюю массу студента по его росту, имеет вид

$$\hat{y} = 76{,}64 + 0{,}51 \cdot (x - 179{,}4).$$

Для построения доверительных интервалов оценок $\hat{A}$ и $\hat{b}$ вычислим:

$$\sigma_x = \sqrt{\overline{x^2} - \left(\overline{x}\right)^2} = \sqrt{\frac{806105}{25} - (179{,}4)^2} = 7{,}736,$$

$$\sigma_y = \sqrt{\overline{y^2} - \left(\overline{y}\right)^2} = \sqrt{\frac{148918}{25} - (76{,}64)^2} = 9{,}1121,$$

$$S_{yx} = \sigma_y \sqrt{\tfrac{n}{n-2}(1 - r_{XY}^2}} = 8{,}5763.$$

Квантиль распределения Стьюдента с числом степеней свободы

$n - 2 = 23$, *порядка* $1 - \dfrac{\alpha}{2}$, равен $t_{1-\alpha/2} = 2{,}069$.

После подстановки в формулы (3.2) и (3.3)получим доверительные интервалы $73{,}0899 < A < 80{,}190$, $0{,}0511 < b < 0{,}9689$.

## 4. МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

При решении инженерных задач часто требуется найти зависимость между случайной величиной $\eta$ и переменными величинами $\xi_1, \xi_2, ..., \xi_m$, значения которых $x_1, x_2, ..., x_m$ задаются заранее при планировании эксперимента. Однако при проведении $n$ экспериментов значения

$$x_{1j}, x_{2j}, ..., x_{mj} \ (j = \overline{1,n})$$

обычно измеряются с некоторыми малыми ошибками (погрешности приборов и др.).

Так как переменные $x_i$ не коррелированы с ошибками измерений, то для получения зависимости между $x_i$ и $y$, где $y$ – значение случайной величины $\eta$, можно использовать обычный метод наименьших квадратов, как при рассмотрении случая линейной регрессии двумерного случайного вектора.

Заметим, что переменная $\eta$ является случайной величиной, так как при проведении эксперимента невозможно учесть все факторы в том числе ошибки измерений, оказывающие влияние на эту переменную,

При исследовании взаимосвязи между случайной величиной $\eta$ и переменными $\xi_1, \xi_2, ..., \xi_m$ обычно рассматриваются следующие вопросы:

– выбор модели регрессии;

– нахождение оценок этих параметров выбранного у и построение доверительных интервалов параметров уравнения по заданному уровню значимости $\alpha$;

– проверка согласованности выбранноймодели с экспериментальными данными и уточнение вида полученного уравнения.

Выбор модели регрессии производится обычно с учетом эмпирических аспектов. Эту задачу подробно рассматривать не будем ввиду ее сложности.

Проанализируем следующие вопросы:

– построение линейного уравнения регрессии и доверительных интервалов для его параметров;

– проверка согласованности полученной модели с экспериментальными данными наиболее простыми способами.

## 4.1 Нахождение оценок параметров линейного уравнения регрессии

Ограничимся построением линейного уравнения регрессии:

$$\eta = \alpha_0^* + \alpha_1^* \xi_1 + \alpha_2^* \xi_2 + ... + \alpha_m^* \xi_m,$$

т.е. найдем наилучшее приближение функции $\eta$ с помощью функции вида

$$\varphi(x_1,...,x_m,a_0,a_1,...,a_m) = a_0 + a_1x_1 + ... + a_mx_m .$$

Пусть в *j*-м эксперименте величины $\xi_1,\xi_2,...,\xi_m$ приняли значения $x_{1j},x_{2j},...,x_{mj}$, а случайная величина $\eta$ – значения $y_j\ (j=\overline{1,n})$.

По методу наименьших квадратов в качестве оценок параметров $\alpha_0^*,\alpha_1^*,...,\alpha_m^*$ принимаем значения $a_0^*,a_1^*,...,a_m^*$, при которых достигает минимума функция

$$\Phi(a_0,a_1,...,a_m) = \sum_{j=1}^{n}(y_j - a_0 - a_1x_{1j} - ...a_mx_{mj})^2 .$$

Согласно условиям экстремума функции $\Phi(a_0,a_1,...,a_m)$ параметры $a_0,a_1,...,a_m$ являются решениями системы

$$\frac{\partial\Phi(a_0,a_1,...,a_m)}{\partial a_i} = 0,\ i = 0,\cdots 1,\cdots ...,\ m$$

Введем следующие обозначения: $X$ – $n\times(m+1)$-мерная матрица наблюдений контролируемых переменных, в которую введен дополнительно первый столбец, состоящий из единиц; $Y$ – n-мерный вектор-столбец наблюдаемых значений случайной величины $\eta$; $A$ – $(m+1)$-мерный столбец параметров $a_i\ (i=0,1,...,m)$.

$$X = \begin{pmatrix} 1 & x_{11} & \dots & x_{m1} \\ 1 & x_{12} & \dots & x_{m2} \\ \dots & \dots & \dots & \dots \\ 1 & x_{1n} & \dots & x_{mn} \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}, \quad A = \begin{pmatrix} a_0 \\ a_1 \\ \dots \\ a_m \end{pmatrix}.$$

В матричных обозначениях эта система примет вид

$$(X^T X)A = X^T Y,$$

здесь $X^T$ – матрица, транспонированная к матрице *X*. Решение этой системы находим по формуле

$$A = (X^T X)^{-1} X^T Y,$$

где $(X^T X)^{-1}$ – матрица, обратная матрице $(X^T X)$. Искомое выборочное уравнение регрессии имеет вид

$$y = a_0 + a_1 x_1 + \dots + a_m x_m .$$

При большом числе переменных задача нахождения вектора решается на ЭВМ с помощью стандартных программ.

### 4.2 Построение доверительных интервалов параметров уравнения репрессии

Построение доверительных интервалов параметров

$$a_i \ (i = 0, 1, \dots, m)$$

легко проводится в случае, если остатки $e_j = y_j - \overline{y_j}$ $(j = \overline{1,n})$ распределены по нормальному закону с параметрами

$$M(e) = 0,\ D(e) = \sigma^2 .$$

Здесь $y_j$ – наблюдаемое в *j*-м эксперименте значение $\eta$; $\overline{y_j}$ – значение у, полученное из уравнения регрессии при подстановке значений $x_{1j}, x_{2j}, ..., x_{mj}$, заданных в *j*-м эксперименте. При малом числе опытов ($n < 50$) применяют приближенные методы проверки нормального распределения остатков. Можно считать, что остатки $e_j$ распределены по нормальному закону, если не менее 95% из них лежат в интервале $(-2S_e, 2S_e)$, где $S_e = \sqrt{\dfrac{\sum_{j=1}^{n} e_j^2}{n-m+1}}$ – оценка дисперсии.

Случай, когда остатки $e_j$ не подчиняются нормальному закону распределения, рассматривать не будем ввиду его сложности.

Для построения $100 \cdot (1-\alpha)\%$-х доверительных интервалов для $\alpha_j^*$ $(j = 0, 1, ..., m)$ по таблице распределения Стьюдента по заданному уровню значимости $\alpha$ и числу степеней свободы $\nu = n - m - 1$ находим критическое значение статистики $t_{1-\frac{\alpha}{2},\nu}$. Доверительные интервалы имеют вид

$$a_i - t_{1-\frac{\alpha}{2},\nu} \cdot S_{a_i} < \alpha_i^* < a_i + t_{1-\frac{\alpha}{2},\nu} \cdot S_{a_i},$$

где $S_{a_i} = S_e \sqrt{a_{i+1,i+1}^{(-1)}}$, $i = 0,1,...,m$.

Здесь $S_{a_i}$ – средние квадратические ошибки коэффициентов $a_i$; $S_e$ – вычисленная выше оценка дисперсии; $a_{i+1,i+1}^{(-1)}$ –

диагональный элемент матрицы $(X^T X)^{-1}$ размера $(m+1)\times(m+1)$, соответствующий переменной $x_i$.

### 4.3 Проверка согласованности модели с экспериментальными данными

Если в уравнении регрессии какая-то из контролируемых переменных $x_i$ незначительно влияет на переменную y, то эту переменную $x_i$ следует исключить из уравнения регрессии.

Выявление статистически незначимых переменных $x_i$ можно рассматривать как проверку гипотезы

$$\mathrm{H}_0 : \alpha_i^* = 0 \,(i = \overline{1,m}),$$

т.е. $\eta$ не коррелировано с $\xi_i$.

Если остатки $e_j$ распределены по нормальному закону, то гипотеза $\mathrm{H}_0$ может быть проверена с помощью статистики $t_i = \dfrac{a_i}{S_{a_i}}, i = \overline{1,m}$. Статистика $t$ имеет распределение Стьюдента с $\nu = n - m - 1$ степенями свободы при условии справедливости гипотезы $\mathrm{H}_0$.

По таблицам распределения Стьюдента находим критическое значение $t_{1-\frac{\alpha}{2},\nu}$, где $\alpha$ – выбранный уровень значимости. Если выполняется условие $|t_i| > t_{1-\frac{\alpha}{2},\nu}$, то нулевая гипотеза отвергается. Следовательно, проверяемый коэффициент уравнения регрессии $\alpha_i^*$ существенно отличается от нуля или, что

то же самое, контролируемая переменная $x_i$ оказывает значимое влияние на переменную *y*.

Если это неравенство не выполняется, то переменная $x_i$ влияет незначительно на переменную *y*. В этом случае уравнение регрессии нужно строить заново, учитывая в нем все переменные, кроме $x_i$. Построение линейной регрессионной модели, у которой все факторы $x_i$ существенно влияют на переменную *y* может закончиться не на первом, а на втором, третьем и т. д. этапе. На каждом из них заново проводится оценка коэффициентов регрессии и анализ влияния каждой переменной.

Если несмещенная оценка среднего квадратического отклонения $S_e$, вычисляемого по приведенной выше формуле, допустима для данной задачи, то считаем, что модель хорошо согласуется с экспериментом. Обычно в практических задачах требуют, чтобы $S_e$ не превышало 10% абсолютной величины наименьшего значения случайной величины $\eta$.

Если $S_e$ велико, то модель регрессии нужно уточнить, т.е. взять большее число опытов и произвести все вычисления заново. Если и это не поможет, то делаем вывод, что выбранная модель плохо согласуется с экспериментом. В этом случае нужно выбирать другой вид зависимости.

**Пример.** Произведено 10 измерений прочности строительного материала у при равном содержании в нем некоторых компонент $\xi_1, \xi_2$. Заданные при проведении эксперимента значения $x_{1j}, x_{2j}$ компонент $\xi_1, \xi_2$ и полученные значения $y_i$ прочности $\eta$ $(j = \overline{1,10})$ сведены в таблицу:

| $x_1$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| $x_2$ | 17,5 | 13,7 | 10,8 | 8,5 | 5,2 | 5 | 4,95 | 4,92 | 4,9 | 4,89 |
| *y* | 13,25 | 15,95 | 17,63 | 18,63 | 19,2 | 19,37 | 19,5 | 19,6 | 19,67 | 19,7 |

Предполагая, что зависимость между величиной $\eta$ и величинами $\xi_1, \xi_2$ линейная, найти оценки $a_0, a_1, a_2$ параметров $\alpha_0^*, \alpha_1^*, \alpha_2^*$ уравнения регрессии; 95% доверительные интервалов параметров $\alpha_0^*, \alpha_1^*, \alpha_2^*$; и проверить согласованность полученной модели регрессии с экспериментом.

*Решение.*

1. Для нахождения оценок $a_0, a_1, a_2$ коэффициентов выборочного уравнения регрессии $y = a_0 + a_1 x_1 + a_2 x_2$ необходимо решить систему алгебраических уравнений. Матрица коэффициентов $X^T X$ этой системы и матрица-столбец свободных членов $X^T Y$ записываются следующим образом:

$$X^T X = \begin{pmatrix} n & \sum x_1 & \sum x_2 \\ \sum x_1 & \sum x_1^2 & \sum x_1 x_2 \\ \sum x_2 & \sum x_1 x_2 & \sum x_2^2 \end{pmatrix}, \; X^T Y = \begin{pmatrix} \sum y \\ \sum x_1 y \\ \sum x_2 y \end{pmatrix}.$$

В нашем случае вычисления дают:

$$X^T X = \begin{pmatrix} 10 & 45 & 80{,}36 \\ 45 & 285 & 253{,}95 \\ 80{,}36 & 253{,}95 & 831{,}501 \end{pmatrix}, \; X^T Y = \begin{pmatrix} 182{,}5 \\ 869{,}61 \\ 1381{,}512 \end{pmatrix}.$$

Решая систему $(X^T X)A = X^T Y$ линейных алгебраических уравнений третьего порядка, находим:

$$A = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 22{,}3634 \\ -0{,}0473 \\ -0{,}4854 \end{pmatrix}.$$

Значит, выборочное уравнение регрессии имеет вид

$$y = 22{,}3634 - 0{,}0473x_1 - 0{,}4854x_2 \qquad (4.1)$$

2. Прежде чем найти доверительные интервалы параметров $\alpha_0^*, \alpha_1^*, \alpha_2^*$ уравнения регрессии, убедимся, что остатки $e_j = y_j - \overline{y_j}, (j = \overline{1,10})$ распределены по нормальному закону. Так как число опытов мало, применим приближенный метод проверки.

Для удобства вычислений составим таблицу:

| $y_j$ | 13,25 | 15,95 | 17,63 | 18,63 | 19,2 | 19,37 | 19,5 | 19,6 | 19,67 | 19,7 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\overline{y_j}$ | 13,869 | 15,666 | 17,027 | 18,096 | 19,65 | 19,7 | 19,677 | 19,644 | 19,607 | 19,564 |
| $e_j$ | -0,619 | 0,234 | 0,603 | 0,534 | -0,45 | -0,33 | -0,177 | -0,044 | 0,063 | 0,136 |

Значения $\overline{y_j}$ вычисляются, исходя из уравнения регрессии (4.1). Используя данные таблицы, находим:

$$\sum_{j=1}^{n} e_j^2 = 1{,}4805 .$$

Средняя квадратическая ошибка

$$S_e = \sqrt{\frac{1{,}4805}{7}} = 0{,}4599 .$$

Так как в интервал $(-2S_e, 2S_e)$ попадают все остатки $e_j$, то можно считать, что остатки распределены по нормальному закону. В этом случае для нахождения доверительных интервалов вычислим диагональные элементы $a_{11}^{(-1)}, a_{22}^{(-1)}, a_{33}^{(-1)}$ матрицы $(X^T X)^{-1}$:

$$det(X^T X) = \begin{vmatrix} 10 & 45 & 80{,}36 \\ 45 & 285 & 253{,}95 \\ 80{,}36 & 253{,}95 & 831{,}501 \end{vmatrix} = 37297{,}5,$$

$$a_{11}^{(-1)} = \frac{1}{det(X^T X)} \begin{vmatrix} 285 & 253{,}95 \\ 253{,}95 & 831{,}501 \end{vmatrix} = 4{,}6246,$$

$$a_{22}^{(-1)} = \frac{1}{det(X^T X)} \begin{vmatrix} 10 & 80{,}36 \\ 80{,}36 & 831{,}501 \end{vmatrix} = 0{,}0498,$$

$$a_{33}^{(-1)} = \frac{1}{det(X^T X)} \begin{vmatrix} 10 & 45 \\ 45 & 285 \end{vmatrix} = 0{,}0221.$$

Далее находим средние квадратические ошибки параметров $a_i$:

$$S_{a_0} = S_e \sqrt{a_{11}^{(-1)}} = 0{,}989,\ S_{a_1} = S_e \sqrt{a_{22}^{(-1)}} =$$

$$= 0{,}1026,\ S_{a_2} = S_e \sqrt{a_{33}^{(-1)}} = 0{,}0684.$$

По заданному уровню значимости $\alpha = 1 - p = 1 - 0{,}95 = 0{,}05$ и числу степенней свободы $\nu = n - m - 1 = 7$ находим критическое значение

$$t_{1-\frac{\alpha}{2},\nu} = t_{0{,}975,7} = 2{,}365.$$

Доверительные интервалы параметров $\alpha_0^*, \alpha_1^*, \alpha_2^*$ имеют вид:

$$22{,}3634 - 2{,}365 \cdot 0{,}989 < \alpha_0^* < 22{,}3634 + 2{,}365 \cdot 0{,}989,$$

$$-0{,}0473-2{,}365\cdot 0{,}1026<\alpha_1^*<-0{,}0473+2{,}365\cdot 0{,}1026,$$

$$-0{,}4854-2{,}365\cdot 0{,}0684<\alpha_2^*<-0{,}4854+2{,}365\cdot 0{,}0684.$$

Окончательно получаем:

$$20{,}024<\alpha_0^*<24{,}702;$$

$$-0{,}289<\alpha_1^*<0{,}195;$$

$$-0{,}647<\alpha_2^*<-324.$$

Данные доверительные интервалы накрывают параметры $\alpha_0^*,\alpha_1^*,\alpha_2^*$ уравнения регрессии с вероятностью

$$p=1-\alpha=0{,}95.$$

3. Так как остатки $e_j=y_j-\overline{y_j},(j=\overline{1{,}10})$ распределены по нормальному закону, то выявление статистически незначимых переменных, можно осуществить с помощью уравнения $t_i=\dfrac{a_i}{S_{a_i}}$:

$$t_1=\frac{0{,}0473}{0{,}1026}=0{,}461,\ t_2=\frac{0{,}4854}{0{,}0684}=7{,}96.$$

Сравним полученные значения $t_1$ и $t_2$, с критическим значением $t_{0{,}975,7}=2{,}365$. Как видно, для полученной статистики $t_2$ превосходит критическое значение. Значит, коэффициент

$a_2$ отличен от нуля с вероятностью 0,95, т.е. переменная $x_2$ оказывает влияние на $y$. Значение $t_1$ меньше критического. Значит, коэффициент $a_1$ незначительно отличается от нуля. Об этом свидетельствует и тот факт, что доверительный интервал для $\alpha_1^*$ накрывает нуль. Следовательно, переменная $x_1$ незначительно влияет на $y$, и ее из выражения регрессии следует исключить.

Уравнение регрессии должно иметь вид $\eta = \beta_0^* + \beta_2^* \xi_2$.

Для построения выборочного уравнения регрессии используем табличные значения $x_2$ и $y$.

Задача ставится теперь так:

1) найти выборочный коэффициент корреляции и оценить его значимость;

2) построить уравнение регрессии $y = b_0^* + b_2^* x_2$ и найти 95%-е доверительные интервалы параметров $\beta_0^*$ и $\beta_2^*$.

*Решение.*

1. Так же, как и ранее, находим:

$$\rho = -0{,}981, \quad \sigma_{x_2} = 4{,}5427, \quad \sigma_y = 2{,}1208, \quad \overline{x_2^2} = 83{,}1501.$$

Для проверки значимости коэффициента корреляции определяем:

$$t_{\text{набл.}} = \frac{0{,}981 \cdot \sqrt{8}}{\sqrt{1-(0{,}981)^2}} \approx 14{,}29.$$

По уровню значимости $\alpha = 0{,}005$ и числу степеней свободы $\nu = n - m - 1 = 10 - 1 - 1 = 8$ находим критическое значение $t_{\gamma;\nu} = t_{0{,}975;8} = 2{,}306$, где $\gamma = 1 - \alpha/2 = 1 - 0{,}05/2 = 0{,}975$. Как

видно, $t_{\text{набл.}} > t_{\gamma;\nu}$. Значит, с вероятностью $p = 0{,}95$ можно утверждать, что $\rho \neq 0$, т.е. случайные величины $\xi_2$ и $\eta$ не являются независимыми.

2. Находим оценки $b_0^*$ и $b_2^*$ коэффициентов $\beta_0^*$ и $\beta_2^*$ соответствующего уравнения регрессии случайной величины $\eta$ на $\xi_2$:

$$b_0^* = 21{,}9302,\ b_2^* = -0{,}458.$$

Выборочное уравнение регрессии имеет вид:

$$y = 21{,}9302 - 0{,}458x_2.$$

Находим 95%-е доверительные интервалы его коэффициентов:

$$21{,}39302 - 2{,}31 \cdot \frac{4{,}5427}{2{,}1208}\sqrt{\frac{1-(0{,}981)^2}{8} \cdot 83{,}1501} <$$

$$< \beta_0^* < 21{,}9302 + 2{,}31 \cdot 1{,}3397,$$

$$-0{,}458 - 2{,}31 \cdot \frac{2{,}1208}{4{,}5427}\sqrt{\frac{1-(-0{,}981)^2}{8}} <$$

$$< \beta_2^* < -0{,}458 + 2{,}31 \cdot 0{,}03202,$$

После преобразований получим:

$$18{,}836 < \beta_0^* < 25{,}025;\ -0{,}5319 < \beta_2^* < -0{,}384.$$

Учебное издание

ВЕРЕМЕНЮК Валентин Валентинович
КОЖУШКО Валерий Васильевич
КРУШЕВСКИЙ Евгений Александрович

# УЧЕБНО-МЕТОДИЧЕСКОЕ ПОСОБИЕ

к лабораторной работе № 2
«Установление зависимости между двумя случайными величинами по результатам их выборок»

Текст публикуется в авторской редакции.
Ответственный за выпуск Е.А. Крушевский

---

Подписано в печать 15.04.2004.
Формат 60x84 1/16. Бумага типографская № 2.
Печать офсетная. Гарнитура Таймс.
Усл.печ.л. 2,6. Уч.-изд.л. 2,0. Тираж 100. Заказ 23.

---

Издатель и полиграфическое исполнение:
Белорусский национальный технический университет.
Лицензия ЛВ № 155 от 30.01.2003. 220013, Минск, проспект Ф.Скорины, 65.