ГЕОМАГНЕТИЗМ И АЭРОНОМИЯ, 2013, том 53, № 6, с. 821–826

УДК 519.241

ПРОВЕРКА СТЕПЕНИ КОРРЕЛЯЦИОННОЙ СВЯЗИ ДВУХ ПЕРЕМЕННЫХ ДЛЯ СЛУЧАЯ ОДНОВРЕМЕННО МАЛЫХ ЗНАЧЕНИЙ КОЭФФИЦИЕНТОВ КОРРЕЛЯЦИИ И РЕГРЕССИИ ПРИ ПОМОЩИ ПЕРЕХОДА В НОВУЮ СИСТЕМУ КООРДИНАТ

© 2013 г. Н. Н. Шефов

Институт физики атмосферы им. А.М. Обухова РАН, г. Москва

e-mail: nikoshefov@yandex.ru

Поступила в редакцию 13.02.2012 г.
После доработки 23.11.2012 г.

При анализе многолетних изменений параметров различных процессов, описывающих состояние геофизических явлений, а также аналогичных процессов, возникает проблема оценки значимости выявляемых трендов, поскольку при ограниченном временном интервале диапазон изменений исследуемой величины оказывается невелик. Это обусловливает малые значения коэффициентов корреляции и регрессии. Такая ситуация обычно трактуется как малая значимость наблюдаемых процессов, что при более детальном анализе не является справедливым выводом.

DOI: 10.7868/S001679401306014X

1. ВВЕДЕНИЕ

При определении коэффициента корреляции по массиву данных (x, y) иногда могут возникать ситуации, когда совокупность точек на графиках расположена достаточно близко к линии регрессии. Тем не менее, она имеет очень малый наклон к оси x, и поэтому коэффициент корреляции r имеет значения, близкие к нулю. На этом основании рассматриваемое временное изменение оценивается как незначимое, в том числе еще и потому, что вычисляемые погрешности коэффициента корреляции σ_r и коэффициента регрессии σ_ρ оказываются сопоставимы со значениями r и ρ соответственно. В действительности речь идет о значениях $r \sim \pm(0.1\text{–}0.3)$,которые обычно считаются имеющими малые значимости. Такая ситуация встречается при анализе многолетних вариаций каких-либо параметров, например, температуры, связанных с климатическими изменениями.

Поэтому для таких ситуаций формальная оценка значимости на основе малого значения коэффициента корреляции не является достаточной, и необходим более детальный анализ.

Целью работы является доказательство существования параметра корреляционной связи двух переменных, а именно, отношения коэффициентов корреляции и регрессии $\frac{r}{\rho}$, определяющего значимость их линейной взаимозависимости, при одновременно малых их значениях.

2. ОСНОВЫ РАСЧЕТА КОЭФФИЦИЕНТОВ КОРРЕЛЯЦИИ И РЕГРЕССИИ

Распределение точек может быть расположено также вдоль оси y с малым наклоном относительно оси ординат. Поэтому и в таком случае нельзя делать вывод об отсутствии связи на основе малого значения коэффициента корреляции. В первую очередь требуется визуальный контроль рисунков с графическим изображением поля корреляции.

Проверить реальный характер степени корреляции можно путем поворота осей координат на 45°, например, вправо относительно исходной системы в случае распределения точек на графике вдоль оси абсцисс, или влево – в случае распределения точек вдоль оси ординат. В этом случае новая система будет иметь координаты (x', y'). В реальной ситуации числовые значения координат могут быть различного порядка. Однако, как будет показано ниже, это может быть учтено в приведенных формулах без непосредственного преобразования координат путем нормировки относительно средних значений $\bar{x}$ и $\bar{y}$ используемых координат.

Соотношение между новыми и старыми координатами при повороте системы координат на 45° определяется формулами [Бронштейн и Семендяев, 1962]

$$x' = \frac{x \mp y}{\sqrt{2}}, \quad y' = \frac{y \pm x}{\sqrt{2}}.$$

Верхние знаки здесь и в последующих формулах соответствуют повороту системы координат вправо, нижние – влево.

Проведенный ниже анализ сделан на основе линейного уравнения регрессии $y = \rho x + b$.

Если исследуемая корреляционная зависимость заведомо имеет нелинейный характер, то, как правило, ее можно привести к линейному виду путем нелинейного преобразования используемых переменных.

Для линейного уравнения регрессии его параметры определяются формулами [Тейлор, 1985]

$$r = \frac{N\Sigma xy - \Sigma x\Sigma y}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}};$$

$$\rho = \frac{N\Sigma xy - \Sigma x\Sigma y}{N\Sigma x^2 - (\Sigma x)^2};\ b = \frac{\Sigma x^2\Sigma y - \Sigma x\Sigma xy}{N\Sigma x^2 - (\Sigma x)^2},$$

где N – число пар значений (x, y).

Если ввести обозначения

$$X = N^2\sigma_x^2 = N\Sigma x^2 - (\Sigma x)^2;$$

$$Y = N^2\sigma_y^2 = N\Sigma y^2 - (\Sigma y)^2;\quad Z = N\Sigma xy - \Sigma x\Sigma y,$$

то искомые параметры будут иметь вид

$$r = \frac{Z}{\sqrt{XY}};\ \rho_x[y(x)] = \frac{Z}{X};\ \rho_y[x(y)] = \frac{Z}{Y},\ \text{так как}$$

$$r^2 = \rho_x\rho_y;\ b = \frac{\Sigma y}{N} - \frac{\Sigma x}{N}\frac{Z}{X} = \frac{\Sigma y}{N} - \rho\frac{\Sigma x}{N}.$$

Отсюда также следует, что $\frac{r}{\rho_x} = \sqrt{\frac{X}{Y}} = \frac{\sigma_x}{\sigma_y}$. В дальнейшем используется только ρ_x.

Погрешности для них определяются формулами

$$\sigma_r = \frac{1 - r^2}{\sqrt{N-1}},\quad \sigma_\rho = \sqrt{\frac{\rho^2 X + Y - 2\rho Z}{(N-2)X}} = \frac{\rho}{r}\sqrt{\frac{1-r^2}{N-2}};$$

$$\sigma_b = \sigma_\rho\sqrt{\frac{\Sigma x^2}{N}};$$

$$\frac{r}{\sigma_r} = \frac{r\sqrt{N-1}}{1-r^2};\quad \frac{\rho}{\sigma_\rho} = r\sqrt{\frac{N-2}{1-r^2}}.$$

Отношение погрешностей равно

$$\frac{\sigma_\rho}{\sigma_r} = \frac{\rho}{r}\sqrt{\frac{N-1}{(N-2)(1-r^2)}}.$$

Отсюда видно, что многие параметры корреляционной связи определяются коэффициентом корреляции r и отношением коэффициентов корреляции и регрессии $\frac{r}{\rho} = \frac{\sigma_x}{\sigma_y}$, которое фактически означает отношение диапазонов изменения переменных x и y, входящих в корреляционную зависимость. Поэтому в реальной ситуации нет необходимости использовать суммы координат, определяющих эти характеристики, а также вычислять среднеквадратические отклонения, поскольку отношение $\frac{r}{\rho}$, вычисляемое при проведении корреляционного анализа, полностью определяет необходимые условия.

В новой системе координат параметры регрессионного уравнения, выраженные через параметры прежней системы, будут иметь вид

$$\Sigma x' = \frac{1}{\sqrt{2}}(\Sigma x \mp \Sigma y);\quad \Sigma y' = \frac{1}{\sqrt{2}}(\Sigma y \pm \Sigma x);$$

$$\Sigma x'y' = \pm\frac{1}{2}(\Sigma x^2 - \Sigma y^2);\quad \Sigma x'\Sigma y' = \pm\frac{1}{2}[(\Sigma x)^2 - (\Sigma y)^2];$$

$$\Sigma x'^2 = \frac{X + Y \mp 2Z + (\Sigma y \mp \Sigma x)^2}{2N} =$$

$$= \frac{X + Y \mp 2r\sqrt{XY} + (\Sigma y \mp \Sigma x)^2}{2N};$$

$$\Sigma y'^2 = \frac{X + Y \pm 2Z + (\Sigma y \pm \Sigma x)^2}{2N} =$$

$$= \frac{X + Y \pm 2r\sqrt{XY} + (\Sigma y \pm \Sigma x)^2}{2N};$$

$$X' = \frac{1}{2}(X + Y \mp 2Z);\quad Y' = \frac{1}{2}(X + Y \pm 2Z);$$

$$Z' = \pm\frac{1}{2}(X - Y).$$

Если необходимо произвести нормировку координат относительно их средних значений, то новые координаты будут

$$\hat{x} = \frac{x}{\bar{x}}\ \text{ и }\ \hat{y} = \frac{y}{\bar{y}},$$

и легко видеть, что

$$\hat{X}' = \frac{1}{2}\left(\frac{X}{\bar{x}^2} + \frac{Y}{\bar{y}^2} \mp \frac{2Z}{\bar{x}\cdot\bar{y}}\right),\quad \hat{Y}' = \frac{1}{2}\left(\frac{X}{\bar{x}^2} + \frac{Y}{\bar{y}^2} \pm \frac{2Z}{\bar{x}\cdot\bar{y}}\right),$$

$$\hat{Z}' = \pm\frac{1}{2}\left(\frac{X}{\bar{x}^2} - \frac{Y}{\bar{y}^2}\right).$$

Поэтому, в действительности, нет необходимости вычислять новые значения (x', y') для новой системы координат и производить заново регрессионный анализ. Фактически, в представленных выше формулах осуществляется простая замена

$$\Sigma x \to \frac{\Sigma x}{\bar{x}} = N,\ \Sigma y \to \frac{\Sigma y}{\bar{y}} = N,$$

$$X \to \frac{X}{\bar{x}^2},\quad Y \to \frac{Y}{\bar{y}^2},\quad Z \to \frac{Z}{\bar{x}\cdot\bar{y}}.$$

В этом случае при нормировке коррелируемых данных на $(\overline{x},\overline{y})\,\frac{r}{\rho}(\overline{x},\overline{y}) = \frac{\overline{y}}{\overline{x}}\frac{r}{\rho}(x,y) = \frac{\overline{y}}{\overline{x}}\frac{\sigma_x}{\sigma_y}$. Если числовые значения используемых координат имеют одинаковые порядки, то нормировка практически не нужна.

Если же сделать нормировку относительно среднеквадратических отклонений σ_x и σ_y, то получается аналогичная замена

$$\Sigma x \to \frac{\Sigma x}{\sigma_x}, \quad \Sigma y \to \frac{\Sigma y}{\sigma_y}, \quad X \to \frac{X}{\sigma_x^2} = \mathrm{N}^2,$$

$$Y \to \frac{Y}{\sigma_y^2} = \mathrm{N}^2, \quad Z \to \frac{Z}{\sigma_x\sigma_y} = r\mathrm{N}^2.$$

Поэтому $\frac{r}{\rho}(\sigma_x,\sigma_y) \equiv 1$. Вследствие этого никакого необходимого вывода сделать нельзя.

3. АНАЛИЗ РЕГРЕССИОННЫХ СООТНОШЕНИЙ

Параметры уравнения регрессии в новой системе координат имеют вид

$$r' = \frac{\pm(X-Y)}{\sqrt{(X+Y)^2 - 4Z^2}} = \frac{\pm(X-Y)}{\sqrt{X^2 + 2(1-2r^2)XY + Y^2}} = \frac{\pm\left(\frac{r^2}{\rho^2} - 1\right)}{\sqrt{\frac{r^4}{\rho^4} + 2(1-2r^2)\frac{r^2}{\rho^2} + 1}};$$

$$\rho' = \frac{\pm(X-Y)}{X \mp 2Z + Y} = \frac{\pm(X-Y)}{X \mp 2\rho X + Y} = \frac{\pm(X-Y)}{X \mp 2r\sqrt{XY} + Y} = \frac{\pm\left(\frac{r^2}{\rho^2} - 1\right)}{\frac{r^2}{\rho^2} \mp 2r\frac{r}{\rho} + 1};$$

$$b' = \frac{1}{\sqrt{2}\mathrm{N}}\left[(\Sigma y \pm \Sigma x) \pm \frac{(\Sigma y \mp \Sigma x)(X-Y)}{X \mp 2Z + Y}\right] = \frac{1}{\sqrt{2}\mathrm{N}}[(\Sigma y \pm \Sigma x) \pm (\Sigma y \mp \Sigma x)\rho'];$$

$$\sigma'_r = \frac{1-r'^2}{\sqrt{\mathrm{N}-1}} = \frac{4(XY - Z^2)}{\sqrt{\mathrm{N}-1}[(X+Y)^2 - 4Z^2]} = \frac{4(1-r^2)XY}{\sqrt{\mathrm{N}-1}[X^2 + 2(1-2r^2)XY + Y^2]} =$$

$$= \sigma_r\frac{4XY}{X^2 + 2(1-2r^2)XY + Y^2} = \frac{4\sigma_r}{\frac{r^4}{\rho^4} + 2(1-2r^2)\frac{\rho^2}{r^2} + 1};$$

$$\sigma'_\rho = \sqrt{\frac{(1 \mp \rho')^2 X + (1 \pm \rho')^2 Y \pm 2(1-\rho'^2)Z}{(\mathrm{N}-2)(X + Y \mp 2Z)}} = \sqrt{\frac{(1 \mp \rho')^2 X + (1 \pm \rho')^2 Y \pm 2r(1-\rho'^2)XY}{(\mathrm{N}-2)(X + Y \mp 2r\sqrt{XY})}} =$$

$$= \sqrt{\frac{(1 \mp \rho')^2 + (1 \pm \rho')^2\frac{\rho^2}{r^2} \pm 2\rho(1-\rho'^2)}{(\mathrm{N}-2)\left(1 + \frac{\rho^2}{r^2} \mp 2\rho\right)}};$$

$$\sigma'_b = \sigma'_\rho\frac{\sqrt{X + Y \mp 2Z + (\Sigma y \mp \Sigma x)^2}}{\sqrt{2}\mathrm{N}} = \sigma'_\rho\frac{\sqrt{X + Y \mp 2r\sqrt{XY} + (\Sigma y \mp \Sigma x)^2}}{\sqrt{2}\mathrm{N}}.$$

Как видно, члены исходных выражений входят в формулы параметров в новой системе координат. Причем основные новые параметры – коэффициенты корреляции и регрессии и их погрешности в новой системе координат вычисляются на основе прежних значений этих параметров в исходной системе координат. Это существенно упрощает возможность применения рассматриваемого подхода к решению поставленной задачи, что легко выполнимо даже при использовании простейших вычислительных средств, не говоря уже о компьютерных возможностях.

Таким образом, представленные соотношения показывают, что критерием достоверности степени корреляции является фактор $\frac{r}{\rho}$ исходной системы данных, или $\frac{\overline{y}}{\overline{x}}\frac{r}{\rho}$. Если $\frac{r}{\rho} \sim 1$ при условии $r \le 0.3$, то степень корреляционной связи мала, т.е. недостоверна, и коэффициент корреляции в новой системе $r' \sim 0.2$. Если же $\frac{r}{\rho} \gg 1$, т.е. $\frac{r}{\rho} \ge 2$, то связь достоверна, так как при $r \le 0.3$ и $\frac{r}{\rho} \sim 2$ в новой системе $r' = 0.6$, и связь можно считать достоверной. Конкретные данные для рассматриваемых случаев представлены в таблице.

Примеры различных видов представления данных и типов полей корреляций представлены на рис. 1*а* и *б*. Они свидетельствуют, что нормировка

Сравнение параметров корреляции в исходной и новой системах координат

Исходный коэф. корреляции	Коэффициент корреляции, (r') в новой системе координат										
	$\frac{r}{\rho}$										
r	1	1.2	1.5	1.7	2	2.2	2.5	3	3.5	4	5
0.000	0.000	0.180	0.385	0.486	0.600	0.658	0.724	0.800	0.849	0.882	0.923
0.100	0.000	0.181	0.386	0.488	0.602	0.659	0.726	0.801	0.850	0.883	0.924
0.200	0.000	0.184	0.391	0.492	0.608	0.665	0.731	0.806	0.854	0.886	0.926
0.300	0.000	0.189	0.400	0.503	0.618	0.675	0.740	0.813	0.860	0.891	0.929
0.400	0.000	0.196	0.414	0.519	0.633	0.690	0.753	0.824	0.869	0.898	0.934

исходных данных на σ_x и σ_y не может обеспечить выполнения поставленной цели контроля степени корреляционной связи рассматриваемых компонентов, поскольку, как это следует из приведенных выше формул, при $\frac{r}{\rho} \equiv 1$ в новой системе $r' \equiv 0$ и $\rho' \equiv 0$.

Обычно критериями значимости считаются случаи, когда $\frac{r}{\sigma_r} \geq 3$ и $\frac{\rho}{\sigma_\rho} \geq 3$. В указанных выше ситуациях малых значений коэффициентов корреляции и регрессии это условие не выполняется. Поэтому целесообразно проверить эти критерии в новой системе координат. Как легко показать,

$$\frac{r'}{\sigma'_r} = \frac{r'\sqrt{\mathrm{N}-1}}{1-r'^2} =$$

$$= \frac{\pm(X-Y)\sqrt{(\mathrm{N}-1)[X^2+2(1-r^2)XY+Y^2]}}{4(1-r^2)XY} =$$

$$= \frac{\pm\left(1-\frac{\rho^2}{r^2}\right)\sqrt{(\mathrm{N}-1)\left[\frac{r^4}{\rho^4}+2(1-r^2)\frac{r^2}{\rho^2}+1\right]}}{4(1-r^2)};$$

$$\frac{\rho'}{\sigma'_\rho} = r'\sqrt{\frac{\mathrm{N}-2}{1-r'^2}} =$$

$$= \rho'\sqrt{\frac{(\mathrm{N}-2)(X \mp 2r\sqrt{XY}+Y)}{(1\mp\rho')^2 X \pm 2r(1-\rho'^2)\sqrt{XY}+(1\pm\rho')^2 Y}} =$$

$$= \rho'\sqrt{\frac{(\mathrm{N}-2)\left(\frac{r}{\rho}\mp 2r+\frac{\rho}{r}\right)}{(1\mp\rho')^2\frac{r}{\rho}\pm 2r(1-\rho'^2)+(1\pm\rho')^2\frac{\rho}{r}}}.$$

Поскольку эти формулы выглядят несколько громоздкими, здесь целесообразно рассмотреть два предельных случая для выяснения свойств типичных ситуаций, которые показаны на рис. 1*а* и *б*. Из рассмотрения приведенных выше формул следует, что, во-первых, если исходный коэффициент корреляции близок к $r \sim 1$, то в новой системе координат он может быть оценен как $r' \sim \frac{\frac{r^2}{\rho^2}-1}{\frac{r^2}{\rho^2}-1} \sim 1$. Во-вторых, если значения коэффициента корреляции малы, т.е. $r \sim 0.1$–0.3, то в новой системе координат его можно оценить по формуле $r' \sim \frac{\frac{r^2}{\rho^2}-1}{\frac{r^2}{\rho^2}+1}$. В этом случае, в свою очередь, могут быть две возможности. Во-первых, если $\frac{r}{\rho} \geq 2$, и в новой системе координат коэффициент корреляции $r' \sim 1$. Таким образом, в действительности, степень связи достаточно значимая (рис. 1*а*).

Во втором случае, если $\frac{r}{\rho} \sim 1$, то и в новой системе координат они будут иметь аналогичное соотношение и коэффициент корреляции в новой системе сохранит свое малое значение $r' \sim 0$ (рис. 1*б*).

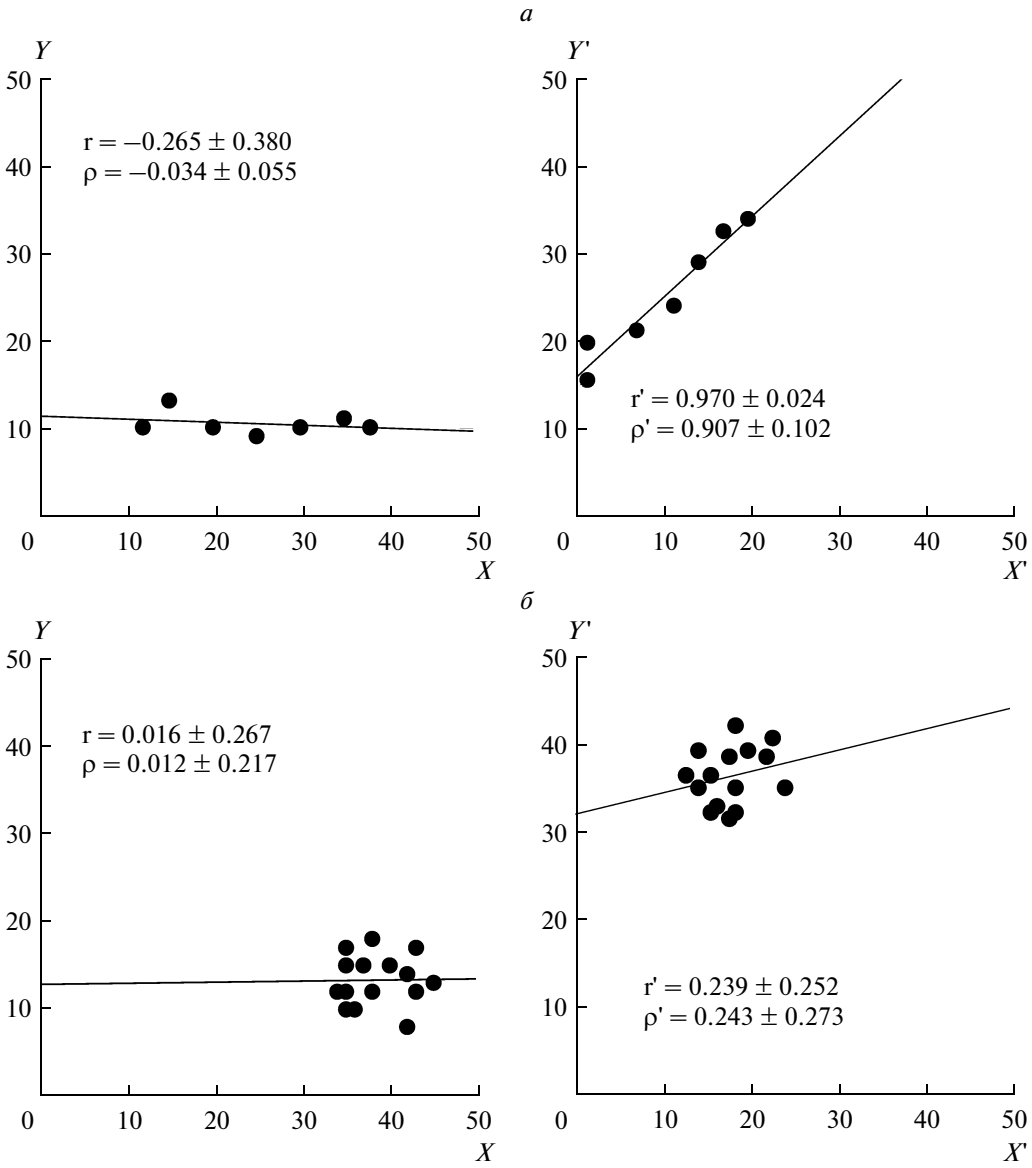

Рис. 1. Варианты типов корреляций: *а* – параметр $\frac{r}{\rho} = 7.780$; *б* – параметр $\frac{r}{\rho} = 1.28$.

Данные Таблицы, представленные на рис. 2, наглядно иллюстрируют зависимость коэффициента корреляции r' от отношения $\frac{r}{\rho}$ для диапазона (0.1–0.3) коэффициента корреляции r.

4. ЗАКЛЮЧЕНИЕ

Таким образом, рассмотренная ситуация позволяет сделать следующий вывод. При коэффициентах корреляции, для которых значимость считается малой, т.е. близких по своим значениям к нулю (0.1–0.3), необходима проверка значения коэффициента корреляции в новой системе координат. Это полезно сделать и визуально. Процедура оценки существенно упрощается тем обстоятельством, что критерием достоверности является отношение коэффициентов корреляции и регрессии $\frac{r}{\rho}$ исходной системы. Как было показано, малые значения коэффициентов регрессии в исходной системе координат означают медленное изменение функции от своего аргумента. Иначе говоря, масштаб относительного изменения аргумента $\frac{\Delta \mathbf{x}}{\Delta \mathbf{y}} = \frac{1}{\rho}$ достаточно велик, при котором происходят соответствующие относительные изменения функции, что не является причиной отсутствия связи между рассматриваемыми переменными. При-

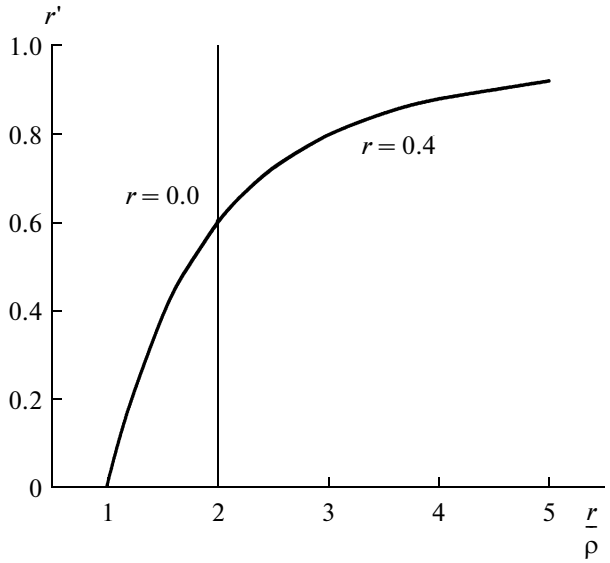

Рис. 2. Зависимость коэффициента корреляции r' в новой системе координат от отношения коэффициентов корреляции и регрессии $\frac{r}{\rho}$ в исходной системе для различных значений коэффициента корреляции $r = 0.0$–0.4.

мер такого рассмотрения уже был представлен на рис. 1*а* и *б*. Из этих данных следует, что в рассматриваемом случае значимость корреляционной связи весьма высока.

Все сказанное не означает, что коэффициент корреляции исходной зависимости равен его значению в новой системе координат. В рассматриваемой ситуации его малого значения он не может быть однозначным критерием значимости изучаемой медленной зависимости, поскольку необходимо более детальное изучение корреляционной связи. Критерием значимости корреляционной зависимости является отношение $\frac{r}{\rho}$.

СПИСОК ЛИТЕРАТУРЫ

– *Бронштейн И.Н., Семендяев К.А.* Справочник по математике. М.: Физматгиз, 608 с. 1962.

– *Тейлор Дж.* Введение в теорию ошибок. М.: Мир, 272 с. 1985.