Основы эконометрики

Общая линейная статистическая модель (ОЛСМ).

Построение ОЛСМ при помощи метода наименьших квадратов.

Линейные и нелинейные модели регрессии.

Корреляция и регрессия.

Использование коэффициента корреляции при построении моделей.

Компьютерная реализация моделирования.

Парная регрессия и корреляция

Парная регрессия – уравнение связи двух переменных y и x:

$$y = \hat{f}(x),$$

где y – зависимая переменная (результативный признак);
x – независимая, объясняющая переменная (признак-фактор).

Линейная регрессия: $y = a + b \cdot x + \varepsilon$.

Нелинейные регрессии

- равносторонняя гипербола $y = a + \frac{b}{x} + \varepsilon$.
- степенная $y = a \cdot x^b \cdot \varepsilon$;
- показательная $y = a \cdot b^x \cdot \varepsilon$;
- экспоненциальная $y = e^{a+b \cdot x} \cdot \varepsilon$.

и другие

Метод наименьших квадратов (МНК)

$$\sum \left(y - \hat{y}_x \right)^2 \rightarrow \min.$$

Для определения параметров линейной модели решается система уравнений:

$$\begin{cases} na + b\sum x = \sum y, \\ a\sum x + b\sum x^2 = \sum yx. \end{cases}$$

МНК может быть использован и при определения параметров нелинейных моделей (см. примеры далее)

Ковариация в построении линейной модели регрессии

$$a = \bar{y} - b \cdot \bar{x}, \qquad b = \frac{\operatorname{cov}(x, y)}{\sigma_x^2} = \frac{\overline{y \cdot x} - \bar{y} \cdot \bar{x}}{\overline{x^2} - \bar{x}^2}.$$

Оценка параметров и качества построенных моделей

Тесноту связи изучаемых явлений оценивает *линейный коэффициент парной корреляции* r_{xy} для линейной регрессии ($-1 \le r_{xy} \le 1$):

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

и *индекс корреляции* ρ_{xy} – для нелинейной регрессии ($0 \le \rho_{xy} \le 1$):

$$\rho_{xy} = \sqrt{1 - \frac{\sigma^2_{ост}}{\sigma^2_y}} = \sqrt{1 - \frac{\sum(y - \hat{y}_x)^2}{\sum(y - \bar{y})^2}}.$$

Оценку качества построенной модели даст коэффициент (индекс) детерминации, а также средняя ошибка аппроксимации.

Средняя ошибка аппроксимации – среднее отклонение расчетных значений от фактических:

$$\overline{A} = \frac{1}{n}\sum\left|\frac{y-\hat{y}}{y}\right| \cdot 100\ \% \ .$$

Допустимый предел значений $\overline{A}$ – не более 8 – 10%.

Средний коэффициент эластичности $\overline{Э}$ показывает, на сколько процентов в среднем по совокупности изменится результат y от своей средней величины при изменении фактора x на 1% от своего среднего значения:

$$\overline{Э} = f'(x)\frac{\overline{x}}{\overline{y}}.$$

Дисперсионный анализ при оценке качества построенной модели

$$\Sigma(y-\bar{y})^2 = \Sigma(\hat{y}_x-\bar{y})^2 + \Sigma(y-\hat{y}_x)^2,$$

где $\Sigma(y-\bar{y})^2$ – общая сумма квадратов отклонений;

$\Sigma(\hat{y}_x-\bar{y})^2$ – сумма квадратов отклонений, обусловленная регрессией («объясненная» или «факторная»);

$\Sigma(y-\hat{y}_x)^2$ – остаточная сумма квадратов отклонений.

Долю дисперсии, объясняемую регрессией, в общей дисперсии результативного признака y характеризует *коэффициент* (*индекс*) *детерминации* R^2:

$$R^2 = \frac{\Sigma(\hat{y}_x-\bar{y})^2}{\Sigma(y-\bar{y})^2}.$$

F-критерий Фишера и t-критерий Стьюдента при оценке качества уравнения регрессии

F-тест – оценивание качества уравнения регрессии – состоит в проверке гипотезы H_0 о *статистической незначимости уравнения регрессии* и *показателя тесноты связи*. Для этого выполняется сравнение фактического $F_{факт}$ и критического (табличного) $F_{табл}$ значений *F-критерия Фишера*. $F_{факт}$ определяется из соотношения значений факторной и остаточной дисперсий, рассчитанных на одну степень свободы:

$$F_{факт} = \frac{\Sigma(\hat{y} - \bar{y})^2 / m}{\Sigma(y - \hat{y})^2 / (n - m - 1)} = \frac{r_{xy}^2}{1 - r_{xy}^2}(n - 2),$$

где n – число единиц совокупности;
m – число параметров при переменных x.

$F_{табл}$ – это максимально возможное значение критерия под влиянием случайных факторов при данных степенях свободы и уровне значимости α. Уровень значимости α – вероятность отвергнуть правильную гипотезу при условии, что она верна. Обычно α принимается равной 0,05 или 0,01.

Если $F_{табл} < F_{факт}$, то H_0 – гипотеза о случайной природе оцениваемых характеристик отклоняется и признается их статистическая значимость и надежность. Если $F_{табл} > F_{факт}$, то гипотеза H_0 не отклоняется и признается статистическая незначимость, ненадежность уравнения регрессии.

Для оценки *статистической значимости коэффициентов регрессии и корреляции* рассчитываются *t-критерий Стьюдента* и *доверительные интервалы* каждого из показателей. Выдвигается гипотеза H_0 о случайной природе показателей, т.е. о незначимом их отличии от нуля. Оценка значимости коэффициентов регрессии и корреляции с помощью *t*-критерия Стьюдента проводится путем сопоставления их значений с величиной случайной ошибки:

$$t_b = \frac{b}{m_b}; \quad t_a = \frac{a}{m_a}; \quad t_r = \frac{r}{m_r}.$$

Случайные ошибки параметров линейной регрессии и коэффициента корреляции определяются по формулам:

$$m_b = \sqrt{\frac{\sum (y - \hat{y}_x)^2 / (n-2)}{\sum (x - \bar{x})^2}} = \sqrt{\frac{S^2_{ост}}{\sum (x - \bar{x})^2}} = \frac{S_{ост}}{\sigma_x \sqrt{n}};$$

$$m_a = \sqrt{\frac{\sum (y - \hat{y}_x)^2}{(n-2)} \cdot \frac{\sum x^2}{n \sum (x - \bar{x})^2}} = \sqrt{S^2_{ост} \frac{\sum x^2}{n^2 \sigma^2_x}} = S_{ост} \frac{\sqrt{\sum x^2}}{n \sigma_x};$$

$$m_{r_{xy}} = \sqrt{\frac{1 - r^2_{xy}}{n-2}}.$$

Сравнивая фактическое и критическое (табличное) значения t-статистики – $t_{табл}$ и $t_{факт}$ – принимаем или отвергаем гипотезу H_0.

Связь между F-критерием Фишера и t-статистикой Стьюдента выражается равенством

$$t_r^2 = t_b^2 = \sqrt{F}\,.$$

Если $t_{табл} < t_{факт}$, то H_0 отклоняется, т.е. a, b и r_{xy} не случайно отличаются от нуля и сформировались под влиянием систематически действующего фактора x. Если $t_{табл} > t_{факт}$, то гипотеза H_0 не отклоняется и признается случайная природа формирования a, b или r_{xy}.

Для расчета доверительного интервала определяем *предельную ошибку* Δ для каждого показателя:

$$\Delta_a = t_{табл} m_a\,, \qquad \Delta_b = t_{табл} m_b.$$

Формулы для расчета *доверительных интервалов* имеют следующий вид:

$$\gamma_a = a \pm \Delta_a; \qquad \gamma_{a_{min}} = a - \Delta_a; \qquad \gamma_{a_{max}} = a + \Delta_a;$$

$$\gamma_b = b \pm \Delta_b; \qquad \gamma_{b_{min}} = b - \Delta_b; \qquad \gamma_{b_{max}} = b + \Delta_b.$$

Прогнозирование с использованием уравнения регрессии. Определение средних стандартных ошибок прогноза

Прогнозное значение y_p определяется путем подстановки в уравнение регрессии $\hat{y}_x = a + b \cdot x$ соответствующего (прогнозного) значения x_p. Вычисляется *средняя стандартная ошибка прогноза* $m_{\hat{y}_p}$:

$$m_{\hat{y}_p} = \sigma_{ост} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_p - \overline{x})^2}{\sum (x - \overline{x})^2}},$$

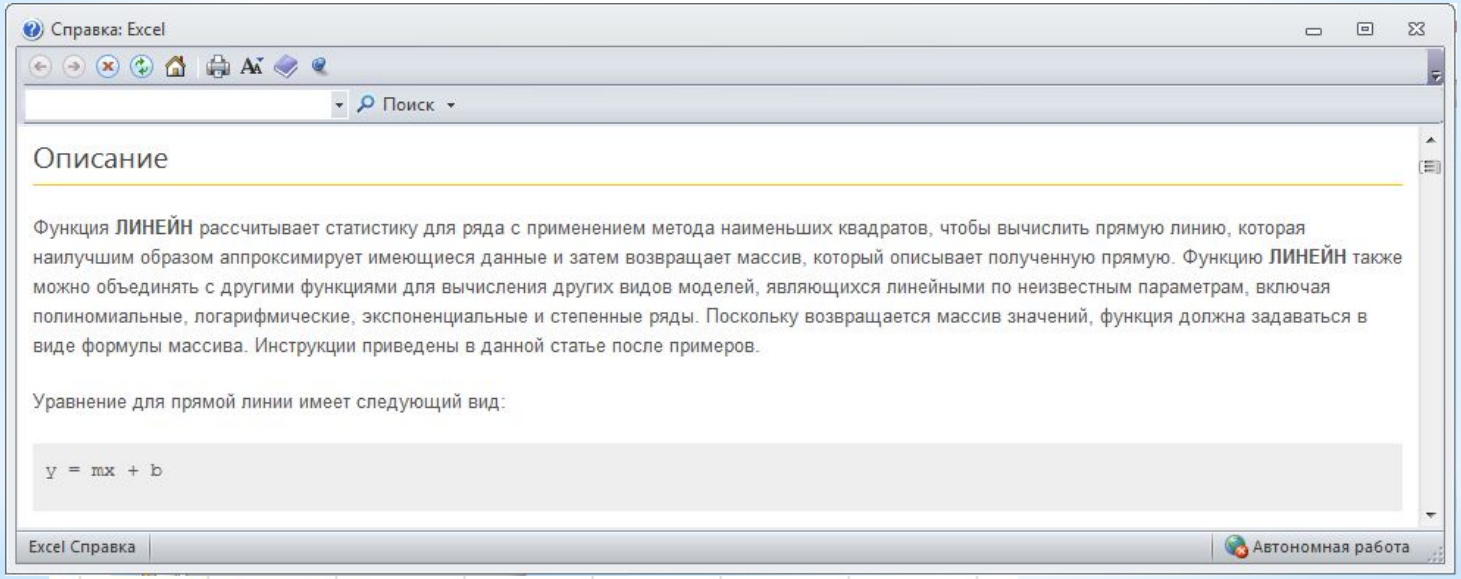
где $\sigma_{ост} = \sqrt{\frac{\Sigma(y - \hat{y})^2}{n - m - 1}}$;

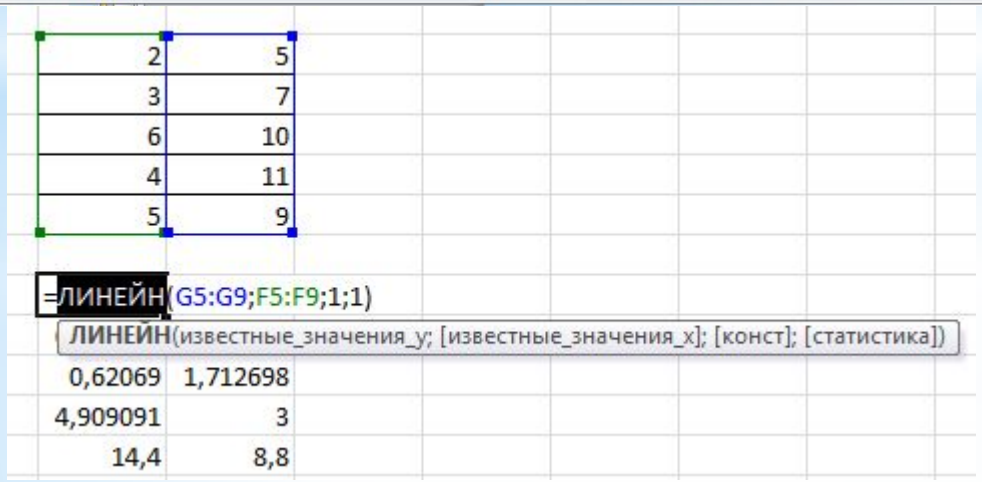
и строится *доверительный интервал прогноза*:

$$\gamma_{\hat{y}_p} = \hat{y}_p \pm \Delta_{\hat{y}_p}; \quad \gamma_{\hat{y}_{p\,min}} = \hat{y}_p - \Delta_{\hat{y}_p}; \quad \gamma_{\hat{y}_{p\,max}} = \hat{y}_p + \Delta_{\hat{y}_p},$$

где $\Delta_{\hat{y}_p} = t_{табл} \cdot m_{\hat{y}_p}$.

Использование встроенной функции ЛИНЕЙН табличного процессора Excel для расчета параметров линейной модели





2	5
3	7
6	10
4	11
5	9
1,2	3,6
0,541603	2,297825
0,62069	1,712698
4,909091	3
14,4	8,8

Линейная модель имеет вид: y=1,2x+3,6

Справка: Excel

Поиск

Величина	Описание
se1,se2,...,sen	Стандартные значения ошибок для коэффициентов m1,m2,...,mn.
seb	Стандартное значение ошибки для постоянной b (seb = #Н/Д, если аргумент ***конст*** имеет значение ЛОЖЬ).
r2	Коэффициент детерминированности. Сравниваются фактические значения y и значения, получаемые из уравнения прямой; по результатам сравнения вычисляется коэффициент детерминированности, нормированный от 0 до 1. Если он равен 1, то имеет место полная корреляция с моделью, т. е. различий между фактическим и оценочным значениями y нет. В противоположном случае, если коэффициент детерминированности равен 0, использовать уравнение регрессии для предсказания значений y не имеет смысла. Дополнительные сведения о способах вычисления r2, см. в подразделе "Замечания" в конце данного раздела.
sey	Стандартная ошибка для оценки y.
F	F-статистика или F-наблюдаемое значение. F-статистика используется для определения того, является ли случайной наблюдаемая взаимосвязь между зависимой и независимой переменными.
df	Степени свободы. Степени свободы полезны для нахождения F-критических значений в статистической таблице. Для определения уровня надежности модели необходимо сравнить значения в таблице с F-статистикой, возвращаемой функцией **ЛИНЕЙН**. Дополнительные сведения о вычислении величины df см. в подразделе "Замечания" в конце данного раздела. Далее в примере 4 показано использование величин F и df.
ssreg	Регрессионная сумма квадратов.
ssresid	Остаточная сумма квадратов. Дополнительные сведения о расчете величин ssreg и ssresid см. в подразделе "Замечания" в конце данного раздела.

На приведенном ниже рисунке показано, в каком порядке возвращается дополнительная регрессионная статистика.

	A	B	C	D	E	F
1	m_n	m_{n-1}	. . .	m_2	m_1	b
2	se_n	se_{n-1}	. . .	se_2	se_1	se_b
3	r_2	se_y				
4	F	d_f				
5	$ss_{рег.}$	$ss_{ост.}$				

Пример решения типовой лабораторной работы

Лабораторная работа №1

Известны значения двух признаков уровня жизни населения по областям и городу Минску за некоторый год (см. таблицу).

Город Минск и области Республики Беларусь	Расходы на покупку продовольственных товаров в общих расходах, %, признак y	Процент населения региона со среднедушевым уровнем располагаемых ресурсов 1.5-3,5 млн. руб. в месяц, %, признак x
г. Минск	68,8	45,1
Минская обл.	61,2	59
Брестская обл.	59,9	57,2
Витебская обл.	56,7	61,8
Гомельская обл.	55	58,8
Гродненская обл.	54,3	47,2
Могилевская обл.	49,3	55,2

Рассчитайте параметры регрессий для линейной, степенной, показательной модели, модели равносторонней гиперболы.
Оцените каждую модель при помощи средней ошибки аппроксимации и критерия Фишера

Модель линейной регрессии:

	y	x	yx	x^2	y^2	$\hat{y}_x$	$y - \hat{y}_x$	A_i
1	68,8	45,1	3102,88	2034,01	4733,44	61,3	7,5	10,9
2	61,2	59,0	3610,80	3481,00	3745,44	56,5	4,7	7,7
3	59,9	57,2	3426,28	3271,84	3588,01	57,1	2,8	4,7
4	56,7	61,8	3504,06	3819,24	3214,89	55,5	1,2	2,1
5	55,0	58,8	3234,00	3457,44	3025,00	56,5	–1,5	2,7
6	54,3	47,2	2562,96	2227,84	2948,49	60,5	–6,2	11,4
7	49,3	55,2	2721,36	3047,04	2430,49	57,8	–8,5	17,2
Итого	405,2	384,3	22162,34	21338,41	23685,76	405,2	0,0	56,7
Среднее значение	57,89	54,90	3166,05	3048,34	3383,68	x	x	8,1
σ	5,74	5,86	x	x	x	x	x	x
σ^2	32,92	34,34	x	x	x	x	x	x

$$b = \frac{\overline{y \cdot x} - \overline{y} \cdot \overline{x}}{\sigma_x^2} = \frac{3166,05 - 57,89 \cdot 54,9}{5,86^2} \approx -0,35,$$

$$a = \overline{y} - b \cdot \overline{x} = 57,89 + 0,35 \cdot 54,9 \approx 76,88.$$

Уравнение регрессии: $\hat{y} = 76,88 - 0,35 \cdot x$

Т.о. при увеличении среднедушевого населения со среднедушевым размером располагаемых ресурсов 1,5-3,5 млн. руб./мес. на 1% доля расходов на продовольственные товары снижается в среднем на 0,35%.

Рассчитаем линейный коэффициент парной корреляции:

$$r_{xy} = b\frac{\sigma_x}{\sigma_y} = -0{,}35 \cdot \frac{5{,}86}{5{,}74} = -0{,}357.$$

Связь умеренная, обратная.

Определим коэффициент детерминации:

$$r_{xy}^2 = (-0{,}35)^2 = 0{,}127.$$

Вариация результата на 12,7% объясняется вариацией фактора x.

Подставляя в уравнение регрессии фактические значения x, определим теоретические (расчетные) значения $\hat{y}_x$. Найдем величину средней ошибки аппроксимации $\overline{A}$:

$$\overline{A} = \frac{1}{n}\sum A_i = \frac{1}{n}\sum \left|y - \hat{y}\right| \cdot 100\ \% = \frac{56\ ,7 \cdot 100\ \%}{7} = 8\ ,1\ \%.$$

В среднем расчетные значения отклоняются от фактических на 8,1%.

Рассчитаем F-критерий:

$$F_{\text{факт}} = \frac{0{,}127}{0{,}873} \cdot 5 = 0{,}7,$$

Полученное значение указывает на необходимость принять гипотезу H_0 о случайной природе выявленной зависимости и статистической незначимости параметров уравнения и показателя тесноты связи.

Модель степенной регрессии:

1б. Построению степенной модели $y = a \cdot x^b$ предшествует процедура линеаризации переменных. В примере линеаризация производится путем логарифмирования обеих частей уравнения:

$$\lg y = \lg a + b \cdot \lg x;$$

$$Y = C + b \cdot X,$$

где $Y = \lg y$, $X = \lg x$, $C = \lg a$.

	Y	X	YX	Y^2	X^2	$\hat{y}_x$	$y - \hat{y}_x$	$(y - \hat{y}_x)^2$	A_i
1	1,8376	1,6542	3,0398	3,3768	2,7364	61,0	7,8	60,8	11,3
2	1,7868	1,7709	3,1642	3,1927	3,1361	56,3	4,9	24,0	8,0
3	1,7774	1,7574	3,1236	3,1592	3,0885	56,8	3,1	9,6	5,2
4	1,7536	1,7910	3,1407	3,0751	3,2077	55,5	1,2	1,4	2,1
5	1,7404	1,7694	3,0795	3,0290	3,1308	56,3	–1,3	1,7	2,4
6	1,7348	1,6739	2,9039	3,0095	2,8019	60,2	–5,9	34,8	10,9
7	1,6928	1,7419	2,9487	2,8656	3,0342	57,4	–8,1	65,6	16,4
Итого	12,3234	12,1587	21,4003	21,7078	21,1355	403,5	1,7	197,9	56,3
Среднее значение	1,7605	1,7370	3,0572	3,1011	3,0194	x	x	28,27	8,0
σ	0,0425	0,0484	x	x	x	x	x	x	x
σ^2	0,0018	0,0023	x	x	x	x	x	x	x

Рассчитаем C и b:

$$b = \frac{\overline{Y \cdot X} - \overline{Y} \cdot \overline{X}}{\sigma_X^2} = \frac{3{,}0572 - 1{,}7605 \cdot 1{,}7370}{0{,}0484^2} \approx -0{,}298;$$

$$C = \overline{Y} - b \cdot \overline{X} = 1{,}7605 + 0{,}298 \cdot 1{,}7370 = 2{,}278.$$

Получим линейное уравнение: $\hat{Y} = 2{,}278 - 0{,}298 \cdot X$.

Выполнив его потенцирование, получим:

$\hat{y} = 10^{2{,}278} \cdot x^{-0{,}298} = 189{,}7 \cdot x^{-0{,}298}$.

Подставляя в данное уравнение фактические значения x, получаем теоретические значения результата $\hat{y}_x$. По ним рассчитаем показатели: тесноты связи – индекс корреляции ρ_{xy} и среднюю ошибку аппроксимации $\overline{A_i}$:

$$\rho_{xy} = \sqrt{1 - \frac{\sum\left(y - \hat{y}_x\right)^2}{\sum(y - \overline{y})^2}} = \sqrt{1 - \frac{28{,}27}{32{,}92}} = 0{,}3758, \qquad \overline{A} = 8{,}0\%.$$

Характеристики степенной модели указывают, что она несколько лучше линейной функции описывает взаимосвязь.

Рассчитываем аналогично предыдущей модели критерий Фишера с сравниваем его с критическим значением.

Модель показательной регрессии:

1в. Построению уравнения показательной кривой $y = a \cdot b^x$ предшествует процедура линеаризации переменных при логарифмировании обеих частей уравнения:

$$\lg y = \lg a + x \cdot \lg b;$$

$$Y = C + B \cdot x,$$

где $Y = \lg y$, $C = \lg a$, $B = \lg b$.

	Y	x	Yx	Y^2	X^2	$\hat{y}_x$	$y - \hat{y}_x$	$(y - \hat{y}_x)^2$	A_i
1	1,8376	45,1	82,8758	3,3768	2034,01	60,7	8,1	65,61	11,8
2	1,7868	59,0	105,4212	3,1927	3481,00	56,4	4,8	23,04	7,8
3	1,7774	57,2	101,6673	3,1592	3271,84	56,9	3,0	9,00	5,0
4	1,7536	61,8	108,3725	3,0751	3819,24	55,5	1,2	1,44	2,1
5	1,7404	58,8	102,3355	3,0290	3457,44	56,4	−1,4	1,96	2,5
6	1,7348	47,2	81,8826	3,0095	2227,84	60,0	−5,7	32,49	10,5
7	1,6928	55,2	93,4426	2,8656	3047,04	57,5	−8,2	67,24	16,6
Итого	12,3234	384,3	675,9974	21,7078	21338,41	403,4	−1,8	200,78	56,3
Среднее значение	1,7605	54,9	96,5711	3,1011	3048,34	x	x	28,68	8,0
σ	0,0425	5,86	x	x	x	x	x	x	x
σ^2	0,0018	34,3396	x	x	x	x	x	x	x

Значения параметров регрессии A и B составили:

$$B = \frac{\overline{Y \cdot x} - \overline{Y} \cdot \overline{x}}{\sigma_x^2} = \frac{96{,}5711 - 1{,}7605 \cdot 54{,}9}{5{,}86^2} \approx -0{,}0023,$$

$$A = \overline{Y} - B \cdot \overline{x} = 1{,}7605 + 0{,}0023 \cdot 54{,}9 = 1{,}887.$$

Получено линейное уравнение: $\hat{Y} = 1{,}887 - 0{,}0023 \cdot x$.

Произведем потенцирование полученного уравнения и запишем его в обычной форме:

$$\hat{y} = 10^{1{,}887} \cdot 10^{-0{,}0023 \cdot x} = 77{,}1 \cdot 0{,}9947^x.$$

Тесноту связи оценим через индекс корреляции ρ_{xy}:

$$\rho_{xy} = \sqrt{1 - \frac{\sum\left(y - \hat{y}_x\right)^2}{\sum (y - \overline{y})^2}} = \sqrt{1 - \frac{28{,}27}{32{,}92}} = 0{,}3589.$$

Связь умеренная.

$\overline{A} = 8{,}0\%$, что говорит о повышенной ошибке аппроксимации, но в допустимых пределах. Показательная функция чуть хуже, чем степенная, она описывает изучаемую зависимость.

Модель регрессии равносторонней гиперболы:

1г. Уравнение равносторонней гиперболы $y = a + b \cdot \frac{1}{x}$ линеаризуется при замене: $z = \frac{1}{x}$. Тогда $y = a + b \cdot z$.

	y	z	yz	z^2	y^2	$\hat{y}_x$	$y - \hat{y}_x$	$(y - \hat{y}_x)^2$	A_i
1	68,8	0,0222	1,5255	0,000492	4733,44	61,8	7,0	49,00	10,2
2	61,2	0,0169	1,0373	0,000287	3745,44	56,3	4,9	24,01	8,0
3	59,9	0,0175	1,0472	0,000306	3588,01	56,9	3,0	9,00	5,0
4	56,7	0,0162	0,9175	0,000262	3214,89	55,5	1,2	1,44	2,1
5	55	0,0170	0,9354	0,000289	3025,00	56,4	−1,4	1,96	2,5
6	54,3	0,0212	1,1504	0,000449	2948,49	60,8	−6,5	42,25	12,0
7	49,3	0,0181	0,8931	0,000328	2430,49	57,5	−8,2	67,24	16,6
Итого	405,2	0,1291	7,5064	0,002413	23685,76	405,2	0,0	194,90	56,5
Среднее значение	57,9	0,0184	1,0723	0,000345	3383,68	x	x	27,84	8,1
σ	5,74	0,002145	x	X	x	x	x	x	x
σ^2	32,9476	0,000005	x	X	x	x	x	x	x

Значения параметров регрессии a и b составили:

$$a = \bar{y} - b \cdot \bar{z} = 57{,}89 - 1051{,}4 \cdot 0{,}0184 = 38{,}5;$$

$$b = \frac{\overline{y \cdot z} - \bar{y} \cdot \bar{z}}{\sigma_z^2} = \frac{1{,}0723 - 57{,}9 \cdot 0{,}0184}{0{,}002145^2} \approx 1051{,}4.$$

Получено уравнение: $\hat{y} = 38{,}5 + 1051{,}4 \cdot \frac{1}{x}$.

Индекс корреляции: $\rho_{xy} = \sqrt{1 - \frac{27{,}84}{32{,}92}} = 0{,}3944.$

$\overline{A} = 8{,}1\%$. По уравнению равносторонней гиперболы получена наибольшая оценка тесноты связи: $\rho_{xy} = 0{,}3944$ (по сравнению с линейной, степенной и показательной регрессиями). $\overline{A}$ остается на допустимом уровне:

2. $$F_{\text{факт}} = \frac{\rho_{yx}^2}{1 - \rho_{yx}^2} \cdot \frac{n - m - 1}{m} = \frac{0{,}1555}{0{,}8445} \cdot 5 = 0{,}92,$$

где $F_{\text{табл}} = 6{,}6 > F_{\text{факт}}$, $\alpha = 0{,}05$.

Следовательно, принимается гипотеза H_0 о статистически незначимых параметрах этого уравнения. Этот результат можно объяснить сравнительно невысокой теснотой выявленной зависимости и небольшим числом наблюдений.