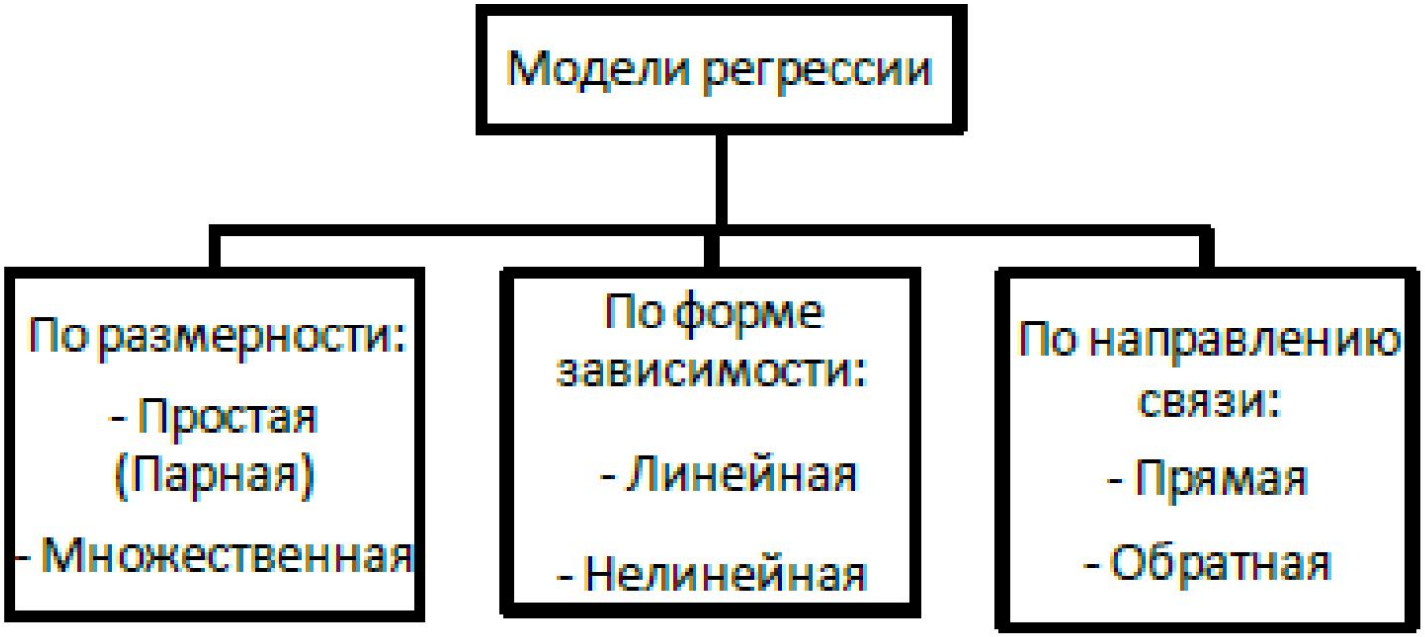
Эконометрика – это научная дисциплина, объединяющая совокупность теоретических результатов, приемов и моделей, предназначенных для того, чтобы на базе экономической теории, экономической и математической статистики придавать конкретное количественное выражение общим закономерностям, установленным экономической теорией.

Рис. 1. Три составляющие эконометрики

При построении эконометрических моделей пользуются инструментарием регрессионного и корреляционного анализа.

Регрессионный анализ предназначен для исследования зависимости изучаемой переменной от различных факторов и отображения их взаимосвязи в форме функции, которая называется *регрессионной моделью*.



Парный регрессионный анализ

Понятие парной регрессии

Предположим, что произведено n наблюдений двух показателей X и Y.

Исходными данными для построения уравнения регрессии служат пары значений $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Парной регрессией называется модель, выражающая зависимость среднего значения зависимой переменной y от одной независимой переменной x

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

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

Знак «^» означает, что между переменными x и y нет строгой функциональной зависимости.

Практически величина y складывается из двух слагаемых:

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

где y – фактическое значение результативного признака;

$\hat{y}$ – теоретическое значение результативного признака, найденное исходя из уравнения регрессии;

ε – случайная величина, возмущение или ошибка модели.

Ее присутствие в модели обусловлено следующими причинами:

1. *Ошибки спецификации модели*, обусловленные не включением важных объясняющих переменных, неправильную функциональную спецификацию модели.

2. *Ошибки измерения*, обусловленные погрешностью сбора и измерения исходных данных.

3. *Ошибки, связанные со случайностью человеческих реакций*. Обусловлено тем, что поведение и непосредственное участие человека в сборе и подготовке данных может внести определенные погрешности.

Спецификация модели

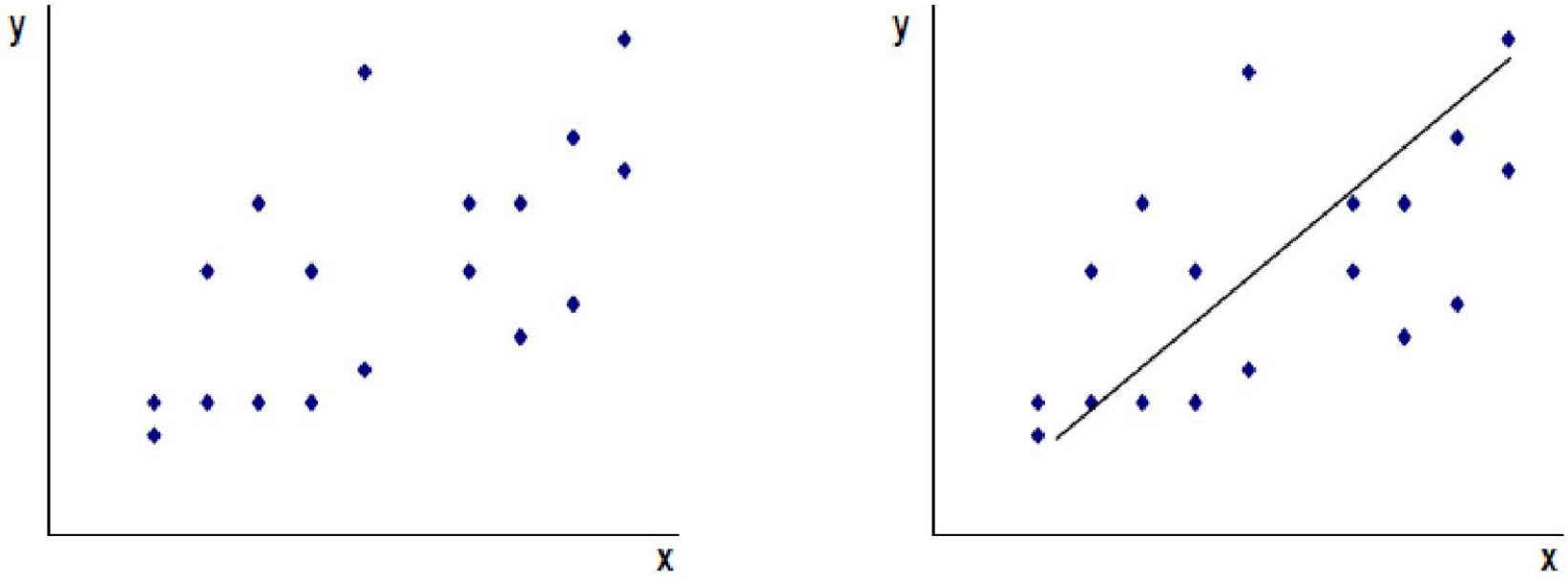
Спецификация модели – формулирование вида модели, исходя из соответствующей теории связи между переменными. Определяется состав переменных и математическая функция для отражения связи между ними.

Для выбора вида аналитической зависимости можно использовать следующие методы:

- *графический* (вид зависимости определяется на основе анализа поля корреляций);
- *аналитический* (на основе качественного анализа изучаемой взаимосвязи);
- *экспериментальный* (построение нескольких моделей различного вида с выбором наилучшей согласно применяемому критерию качества).

При изучении зависимости между двумя признаками графический метод подбора вида уравнения регрессии достаточно нагляден. Он основан на поле корреляции.

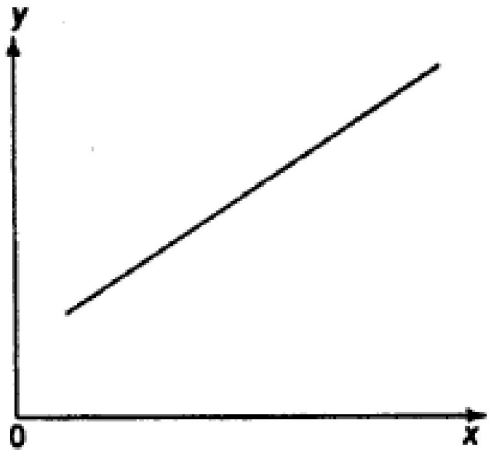
Корреляционное поле



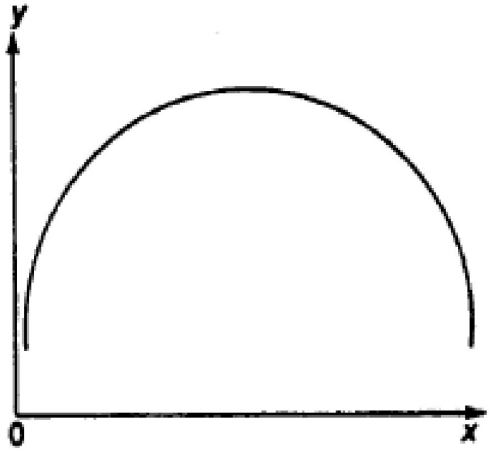
Визуальный анализ поля корреляций позволяет определить форму кривой регрессии, ее особенности.

Зная типичный вид графиков различных функций можно подобрать соответствующую аналитическую зависимость.

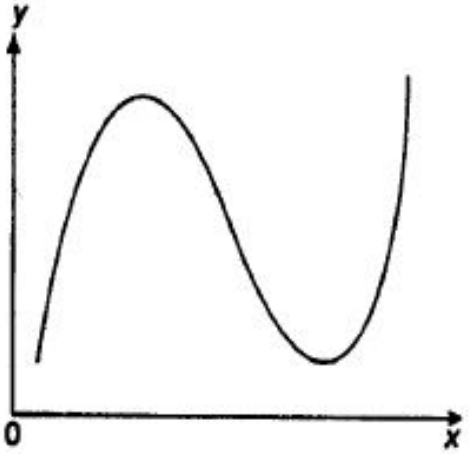
Основные типы кривых



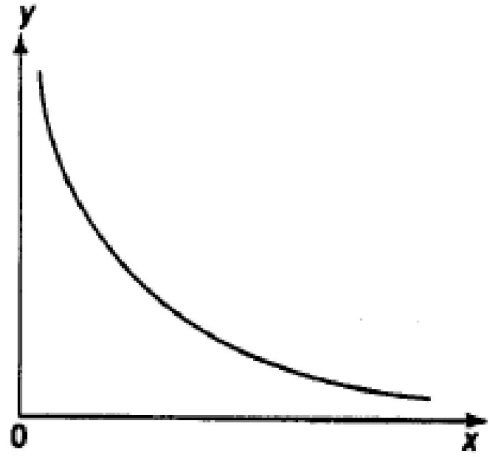
$\hat{y}_x = a + b \cdot x$



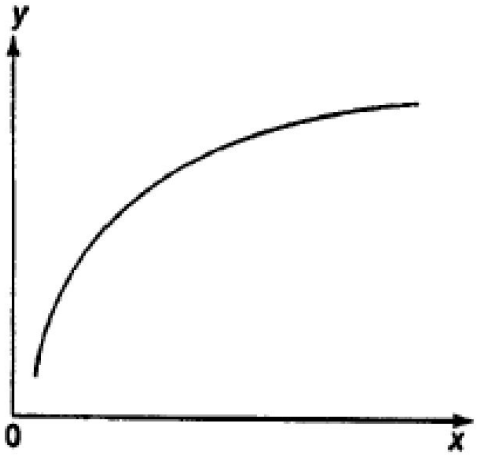
$\hat{y}_x = a + b \cdot x + c \cdot x^2$



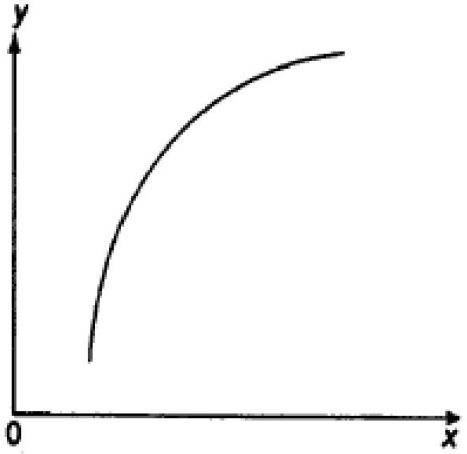
$\hat{y}_x = a + b \cdot x + c \cdot x^2 + d \cdot x^3$



$\hat{y}_x = a + b/x$



$\hat{y}_x = a \cdot x^b$



$\hat{y}_x = a \cdot b^x$

Рассмотрим простейшую модель парной регрессии – *линейную регрессию*.

Линейная парная регрессия описывается уравнением:

$$\hat{y} = a + b \cdot x \text{ или } y = a + b \cdot x + \varepsilon,$$

согласно которому изменение Δy переменной y прямо пропорционально изменению Δx переменной x ($\Delta y = b \cdot \Delta x$).

Построение линейной регрессии сводится к оценке ее параметров *a* и *b*. Классический подход к оцениванию параметров линейной регрессии основан на *методе наименьших квадратов* (МНК).

Согласно МНК, выбираются такие значения параметров *a* и *b*, при которых сумма квадратов отклонений фактических значений результативного признака y_i от теоретических значений $\hat{y}_i = f(x_i)$ (при тех же значениях фактора x_i) минимальна, т. е.

$$S = \sum (y_i - \hat{y}_i)^2 \to \min$$

Система нормальных уравнений метода наименьших квадратов

$$\begin{cases} na + b\sum x_i = \sum y_i; \\ a\sum x_i + b\sum x_i^2 = \sum x_i y_i. \end{cases}$$

Откуда следуют следующие выражения для определения параметров a и b

$$b = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} \qquad a = \overline{y} - b\overline{x}$$

$$\overline{x} = \frac{1}{n}\sum x \qquad \overline{y} = \frac{1}{n}\sum y \qquad \overline{y \cdot x} = \frac{1}{n}\sum y \cdot x \qquad \overline{x^2} = \frac{1}{n}\sum x^2$$

Коэффициент b при факторной переменной x называется *коэффициентом регрессии* и показывает, на сколько изменится в среднем величина y при изменении фактора x на единицу.

Например, допустим, что зависимость между затратами y (тыс. руб.) и объемом выпуска продукции x (ед.) описывается соотношением

$$\hat{y} = 35000+0{,}58\cdot x.$$

В этом случае увеличение объема выпуска продукции на 1 единицу потребует дополнительных затрат в среднем в размере 0,58 тыс. руб. (или 580 рублей).

Параметр a может не иметь экономического содержания.

Линейный коэффициент корреляции r_{xy}:

$$r_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n\sigma_x\sigma_y} = \frac{\overline{yx} - \bar{y}\cdot\bar{x}}{\sigma_x\sigma_y}$$

$$-1 \le r_{xy} \le 1$$

Для качественной оценки тесноты связи можно использовать следующую классификацию:

$0 \le r_{xy} \le 0{,}3$ – очень слабая связь;
$0{,}3 \le r_{xy} \le 0{,}5$ – слабая связь;
$0{,}5 \le r_{xy} \le 0{,}7$ – умеренная связь;
$0{,}7 \le r_{xy} \le 0{,}9$ – тесная связь;
$0{,}9 \le r_{xy} \le 0{,}99$ – очень тесная.

Коэффициент линейной парной корреляции может быть определен через коэффициент регрессии b:

$$r_{xy} = b\frac{\sigma_x}{\sigma_y}$$

Для оценки качества подбора линейной функции рассчитывается квадрат линейного коэффициента корреляции, называемый *коэффициентом детерминации*. Коэффициент детерминации характеризует долю дисперсии результативного признака *y*, объясняемую регрессией, в общей дисперсии результативного признака:

$$r_{xy}^2 = \frac{\sigma_{факт}^2}{\sigma_y^2} = 1 - \frac{\sigma_{ост}^2}{\sigma_y^2}$$

$\sigma_y^2 = \frac{1}{n}\sum_i (y_i - \bar{y})^2$ - общая дисперсия

$\sigma_{факт}^2 = \frac{1}{n}\sum_i (\hat{y}_i - \bar{y})^2$ - факторная дисперсия

$\sigma_{ост}^2 = \frac{1}{n}\sum_i (y_i - \hat{y}_i)^2$ - остаточная дисперсия

Соответственно величина $1 - r_{xy}^2$ характеризует долю дисперсии *y*, вызванную влиянием остальных, не учтенных в модели, факторов.

Чтобы иметь общее суждение о качестве модели из относительных отклонений по каждому наблюдению, определяют *среднюю ошибку аппроксимации*:

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

Средняя ошибка аппроксимации не должна превышать 8–10%.

Оценка значимости уравнения регрессии в целом производится на основе *F*-критерия Фишера. Для парной линейной регрессии он рассчитывается по следующей формуле:

$$F = \frac{r_{xy}^2}{1 - r_{xy}^2}(n-2)$$

Фактическое значение *F*-критерия Фишера сравнивается с табличным значением *F*табл (α; *k*1; *k*2) при уровне значимости α и степенях свободы *k*1 = *m* и *k*2 = *n* – *m* – 1 (*n* – число наблюдений, *m* – число параметров при переменной *x*). Для парной линейной регрессии *m* = 1, поэтому *k*2 = *n* – 2. При этом, если фактическое значение *F*-критерия больше табличного, то признается статистическая значимость уравнения в целом.

Для оценки статистической значимости отдельных параметров уравнения рассчитываются *t*-критерии Стьюдента.

Выдвигается гипотеза H_0 о случайной природе показателей, т.е. о незначимом их отличии от нуля. Рассчитываются фактические значения *t*-критерия:

	a	*b*	*r*
Стандартная ошибка	$m_a = S_{ост} \frac{\sqrt{\sum x^2}}{\sigma_x n}$	$m_b = \frac{S_{ост}}{\sigma_x \sqrt{n}}$	$m_r = \sqrt{\frac{1 - r_{xy}^2}{n - 2}}$
t-критерий	$t_a = \frac{a}{m_a}$	$t_b = \frac{b}{m_b}$	$t_r = \frac{r}{m_r}$
Доверительный интервал	$a \pm t_{табл} \cdot m_a$	$b \pm t_{табл} \cdot m_b$	

Фактические значения *t*-статистики сравниваются с табличным значением *tтаб*(α, n - 2) при определенном уровне значимости α и числе степеней свободы (n – 2).

Если $t_{таб} < t_{факт}$, то H_0 отклоняется, т.е. параметр (*a* или *b*) не случайно отличаются от нуля и сформировался под влиянием систематически действующего фактора *x* (параметр значим).

Если $t_{таб} > t_{факт}$, то H_0 принимается и признается случайная природа формирования параметра (параметр не значим).

Существует связь между *t*-критерием Стьюдента и *F*-критерием Фишера:

$$t_r^2 = t_b^2 = F$$

Т.о., проверка гипотез о значимости коэффициента регрессии и корреляции равносильна проверке гипотезы о существенности линейного уравнения регрессии.

Прогнозное значение y_p определяется путем подстановки в уравнение регрессии

$$\hat{y}_x = a + b \cdot x$$

соответствующего прогнозного значения $x_{пр}$.

Такой прогноз называется точечным. Однако точечный прогноз явно нереален, поэтому вычисляется стандартная ошибка прогноза m_{yp}

$$m_{y_{пр}} = S_{ост} \cdot \sqrt{1 + \frac{1}{n} + \frac{\left(x_{пр} - \bar{x}\right)^2}{n \cdot \sigma_x^2}}, \quad S_{ост} = \sqrt{\frac{\sum \left(y - \hat{y}_x\right)^2}{n - m - 1}}$$

и строится доверительный интервал прогноза

$$y_{пр} - t_{табл} \cdot m_{y_{пр}} \le y_p^* \le y_{пр} + t_{табл} \cdot m_{y_{пр}}$$