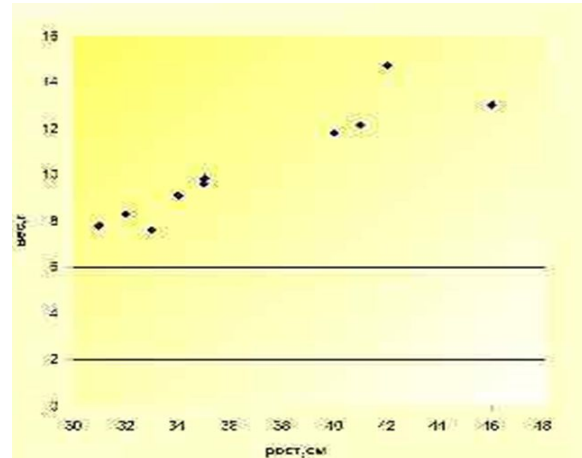
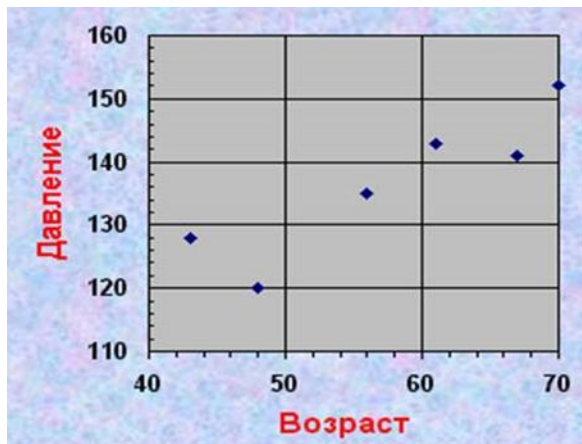
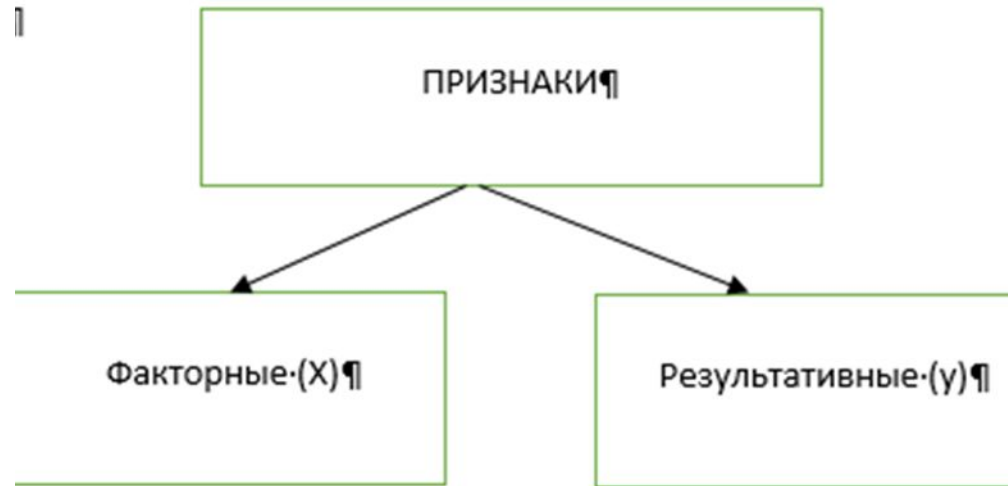


# Корреляционно-регрессионный анализ

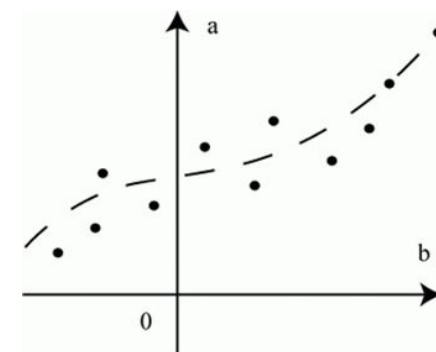
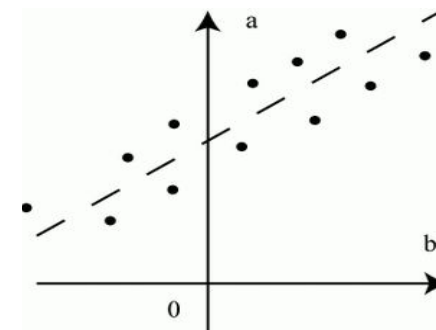
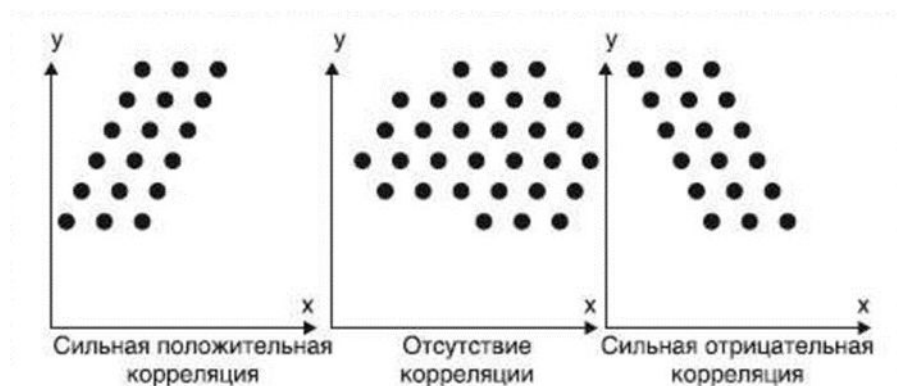
Максименко И.И.

к.э.н., доцент каф. мировой и региональной экономики,  
экономической теории

# Виды признаков



# Виды взаимосвязи признаков

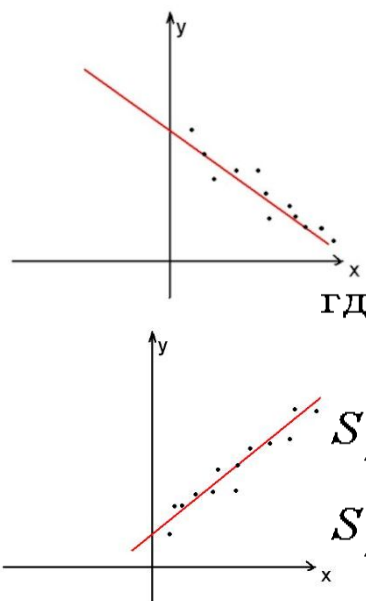


$$r_{xy} = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} = \frac{\overline{xy} - \bar{x}\bar{y}}{S_x \cdot S_y},$$

где  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ ,  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ ,  $\overline{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i$  – выборочные средние,

$S_y^2 = \frac{1}{n} \left( \sum (y_i - \bar{y})^2 \right)$  и  $S_x^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$  – выборочные дисперсии,

$S_y$ ,  $S_x$  – среднеквадратические отклонения.



## При использовании корреляционно-регрессионного анализа необходимо соблюдать следующие требования.

1. Совокупность исследуемых исходных данных должна быть однородной и математически описываться непрерывными функциями.
2. Все факторные признаки должны иметь количественное (цифровое) выражение.
3. Необходимо наличие массовости значений изучаемых показателей.
4. Причинно-следственные связи между явлениями и процессами могут быть описаны линейной или приводимой к линейной формой зависимости.
5. Не должно быть количественных ограничений на параметры модели связи.
6. Необходимо обеспечить постоянство территориальной и временной структуры изучаемой совокупности.

Признаки по их значению делятся на 2 класса.

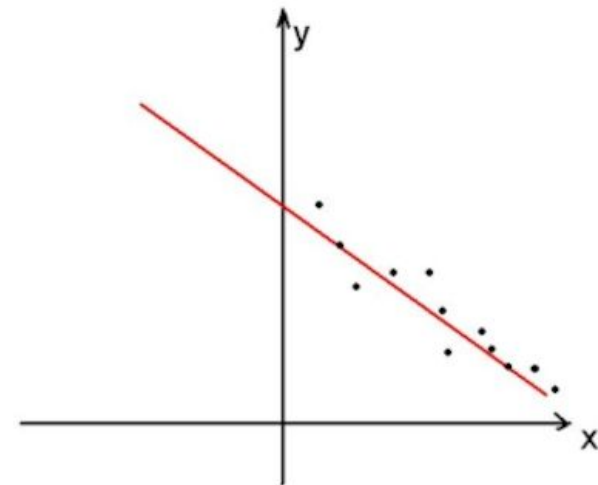
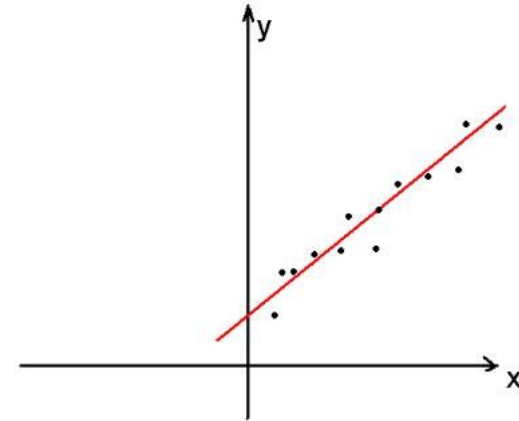
- 1. Результативные признаки* – признаки, изменяющиеся под действием других связанных с ними признаков.
- 2. Факторные* – признаки, обуславливающие изменения результативных признаков.

# Задачи корреляционного анализа:

- выделение важнейших факторов, которые влияют на результативный признак;
- измерение тесноты связи между факторами;
- выявление неизвестных причин связей;
- оценка факторов, оказывающих максимальное влияние на результат.

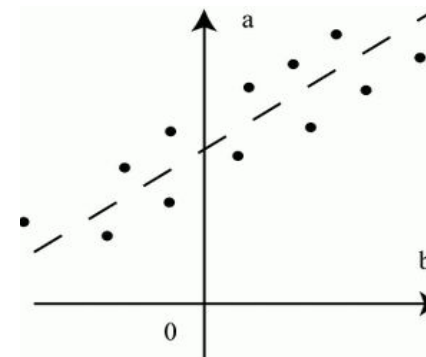
## *Зависимость по направлению связи:*

- Положительная (прямая) – с увеличением (уменьш) одного признака в основном увелич. (уменьш) значения другого.
- Отрицательная (обратная) – с увеличением (уменьш) одного признака в основном уменьшаются (увеличив) значения другого.

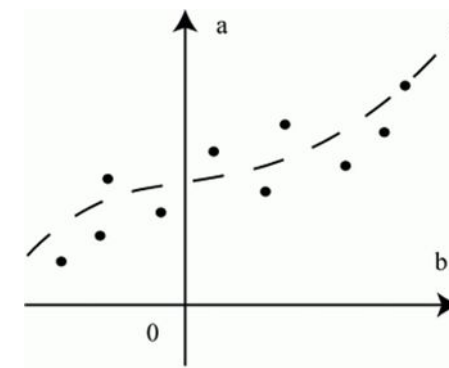


## *Относительно своей аналитической формы:*

- Линейная – между признаками в среднем проявляются линейные соотношения.



- Нелинейная – выражается нелинейной функцией, а переменные связаны между собой в среднем нелинейно.





## *Виды зависимостей:*

1. Парная корреляция - связь между двумя признаками (результативным и факторным).
2. Частная корреляция - зависимость между результативным и одним из факторных признаков при фиксированном значении других факторных признаков.
3. Множественная корреляция - зависимость результативного и двух или более факторных признаков, включенных в исследование.

## линейный коэффициент корреляции $r_{yx}$

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

где  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ ,  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ ,  $\overline{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i$  – выборочные средние,

$S_y^2 = \frac{1}{n} \left( \sum (y_i - \bar{y})^2 \right)$  и  $S_x^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$  – выборочные дисперсии,

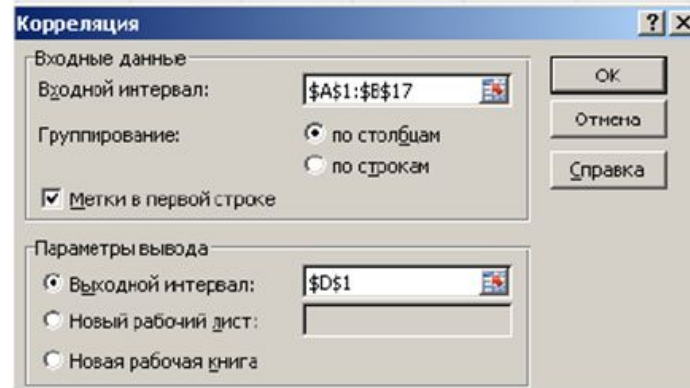
$S_y$ ,  $S_x$  – среднеквадратические отклонения.

Линейный коэффициент корреляции принимает значения от  $-1$  до  $+1$ . Если  $|r_{yx}| \geq 0,7$ , то связь считается сильной. Если  $|r_{yx}| < 0,7$ , то связь считается слабой. Этот коэффициент дает объективную оценку лишь при линейной зависимости.

# Порядок анализа в MS EXCEL

Откройте модуль «Анализ данных», выберите опцию «Корреляция», после чего щелкните мышкой «ОК».

В появившемся окне выполните операции и установки, как показано на рисунке:



Стартовая панель (ваши значения в ячейке **входной** и **выходной** интервал могут отличаться, **метки в первой строке** ставятся при условии, что входной интервал включает заголовки таблицы X и Y).

Щелкнете мышкой «ОК». Результат обработки появится в указанном поле (выходной интервал \$E\$1).

Результат обработки:

	X	Y
X	1	
Y	0,85	1

В полученной таблице нас интересует значение в ячейке на пересечении X и Y — 0,85. Это и есть значение *коэффициента корреляции*.

Пример 1.

Определите связь между ВРП Пермского края (млн.руб.), объемом туристического потока (тыс.чел.), количеством несанкционированных свалок (шт.) и числом высаженных деревьев и кустарников на территории Пермского края.

год	врп	деревья	туристы	свалки
2013	880264,4	756274	609	1074
2014	974192,9	177410	549	986
2015	1063780	380999	642	1273
2016	1095969	484679	662	1432
2017	1191441	1118712	663	1492
2018	1318473	965000	738	1093

	врп	деревья	туристы	свалки
врп	1			
деревья	0,562923	1		
туристы	0,858166	0,679717	1	
свалки	0,350364	0,356803	0,3735	1

1. Линейная связь между ВРП Пермского края и объемом туристического потока сильная ( $r=0,858166$ ).
2. Линейная связь между ВРП Пермского края и количеством несанкционированных свалок практически отсутствует ( $r=0,350364$ ).
3. Линейная связь между объемом туристического потока и числом высаженных деревьев и кустарников на территории Пермского края сильная ( $r=0,679717$ ).

Задача.

Определите связь между себестоимостью добычи нефти в РФ, средневзвешенным курсом доллара США, уровне инфляции в России, ценой на нефть марки Brent.

год	с/с	курс· доллар	инфляция	цена·на· нефть
2013	7723,58	31,85	6,45	108,8
2014	8603,35	38,42	11,36	98,9
2015	9596,43	60,96	12,91	52,4
2016	9133,73	67,03	5,4	44
2017	11184,28	58,35	2,5	49,97
2018	12400,14	62,71	4,3	71,38