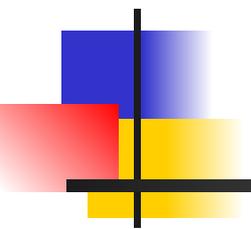


# КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫЙ АНАЛИЗ

- 
- взаимосвязанные показатели
  - часто связь усложняется наложением действием других причин (факторов)
  - Изучить, насколько изменение одного показателя зависит от изменения другого (или нескольких),  
- одна из важнейших задач Статистики

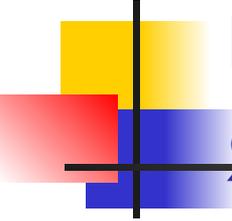


# функциональные и корреляционные

---

- каждому значению одной переменной **строго** соответствует определенное значение другой переменной
- **одному** значению переменной ( $x$ ) может соответствовать **множество значений** другой переменной ( $y$ )

- 
- 
- Наиболее простым случаем корреляционной зависимости является *парная* корреляция, т.е. зависимость между двумя признаками (результативным и одним из факторных).



# Основными задачами при изучении корреляционных зависимостей являются:

---

1) отыскание формы связи в виде математической формулы, выражающей эту зависимость

$y$  от  $x$ ;

2) измерение тесноты такой зависимости

# Возможны различные формы СВЯЗИ:

прямолинейная:  $\bar{y}_x = a_0 + a_1x;$

криволинейная в виде:

а) параболы второго порядка (или высших порядков);

$$\bar{y}_x = a_0 + a_1x + a_2x^2 \text{ и т.д.}$$

б) гиперболы  $\bar{y}_x = a_0 + \frac{a_1}{x};$

в) показательной функции  $\bar{y}_x = a_0a_1^x$  и т.д.

# метод наименьших квадратов (МНК)

---

$$\sum (y - \bar{y}_x)^2 \rightarrow \min$$

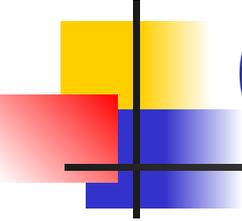
$$(\bar{y}_x = a_0 + a_1 x) \text{ как } \sum (y - a_0 - a_1 x)^2 \rightarrow \min,$$

Линейный коэффициент  
корреляции можно выразить  
формулами:

---

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \sigma_y}$$

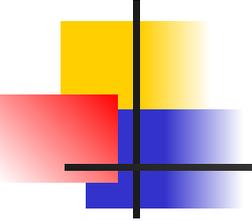
$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{n \sigma_x \sigma_y}$$



# Оценка значимости (существенности)

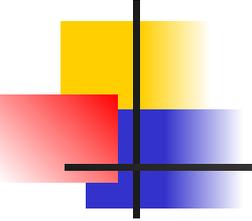
---

- линейного коэффициента корреляций основана на сопоставлении значения  $r$  с его средней квадратической ошибкой ( $\sigma_r$ ).



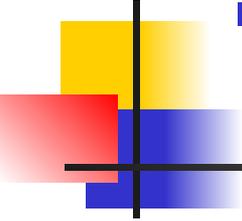
Средняя ошибка коэффициента  
корреляции при  $n > 50$   
рассчитывается приближенно  
по формуле

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

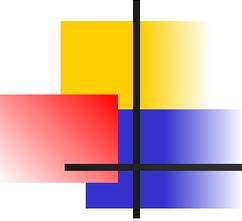
- 
- Если при этом коэффициент корреляции  $r$  превышает свою среднюю ошибку  $\sigma_r$  больше чем в 3 раза, т.е. если

$$\frac{|r|}{\sigma_r} > 3$$

то он считается Значимым, а связь реальной.

- 
- При  $n < 30$  значимость коэффициента корреляции проверяется на основе критерия Стьюдента. Для этого рассчитывается фактическое (расчетное) значение критерия

$$t_{\text{факт}} = \frac{|r| \sqrt{n-2}}{\sqrt{1-r^2}}$$

- 
- 
- Если  $t_{\text{факт}} > t_{\text{табл}}$  коэффициент корреляции  $r$  считается значимым, а связь — реальной.
  - Если  $t_{\text{факт}} < t_{\text{табл}}$ , то считается, что связь между  $x$  и  $y$  отсутствует и значение  $r$ , отличное от нуля, получено случайно.



# ИТАК:

---

На **первом шаге**

регрессионного анализа

идентифицируют переменные ,

$X_1, \dots, X_p$

от которых зависит ,

$Y$

т.е. определяют те

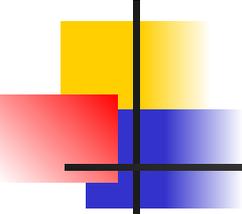
существенные факторы,

которые воздействуют на этот

показатель. Символически этот

факт записывается так:

$$Y = f(X_1, \dots, X_p)$$

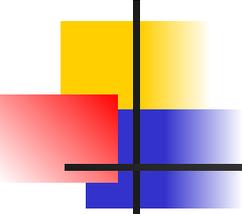


---

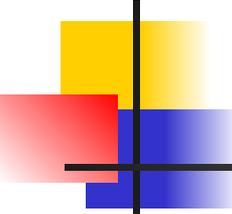
На **втором шаге** регрессионного анализа требуется *спецификация* формы связи между

$$Y \quad X_1, \dots, X_p$$

т.е. определение вида функции .  $f$   
Ориентиром для определения вида зависимости являются содержание решаемой задачи, результаты наблюдений за поведением показателя относительно изменения факторов на основе статистических данных.

- 
- 
- Задача **третьего шага** регрессионного анализа заключается в определении конкретных числовых значений параметров на основе статистических данных о наблюдениях значений  $Y$  и  $X_1, \dots, X_p$

- На практике регрессия чаще всего ищется в виде линейной функции: (линейная регрессия), наилучшим образом приближающей искомую кривую. Делается это с помощью метода наименьших квадратов.

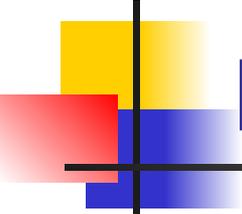


## наиболее важные параметры регрессионной модели

---

**Multiple R** - коэффициент множественной корреляции, который характеризует тесноту линейной связи между зависимой и всеми независимыми переменными. Может принимать значения от 0 до 1.

**$R^2$  - коэффициент детерминации.** Численно выражает долю вариации зависимой переменной, объясненную с помощью регрессионного уравнения. Чем больше  $R^2$ , тем большую долю вариации объясняют переменные, включенные в модель. Например  $R^2=0,76$  - значит уравнение описывает 76% общей дисперсии модели.



# наиболее важные параметры регрессионной модели

---

При поиске лучшей регрессионной модели следует руководствоваться следующими наиболее общими требованиями (Дрейпер, Смит, 1981):

- Регрессионная модель должна объяснять не менее 80% вариации зависимой переменной, т.е.  $R^2 = 0.8$ .
- Стандартная ошибка оценки зависимой переменной по уравнению должна составлять не более 5% среднего значения зависимой переменной;
- Коэффициенты уравнения регрессии и его свободный член должны быть значимы на 5%-ом уровне.
- Остатки от регрессии должны быть без заметной автокорреляции ( $r < 0,30$ ), нормально распределены и без систематической составляющей.

# Проверка значимости модели

Часто F-критерий можно рассчитать через коэффициент корреляции  $r$ :

$$F = \frac{r^2}{1 - r^2} \cdot \frac{n - m}{m - 1}$$

$m$  – число параметров в уравнении регрессии

Расчетное  $F$  сопоставляется с табличным, определяемым по таблице для числа степеней свободы  $\nu_1 = m - 1$  и  $\nu_2 = n - m$  при заданном уровне значимости (например  $\alpha = 0,05$ ).

**Если  $F_{расч} > F_{табл}$ , то уравнение считается значимым.**