**Тема лекции:**

# Статистическое изучение взаимосвязи показателей таможенной статистики

# Учебные вопросы:

1. Понятие статистической зависимости. Постановка задачи корреляционно-регрессионного анализа.
2. Методы выявления взаимосвязи. Количественная оценка тесноты связи между показателями таможенной статистики.
3. Модель взаимосвязи между показателями таможенной статистики.

# Два класса признаков

- **Факторные (Х)**
- **Результативные (У)**

# *Виды связей*

- *Функциональная*
- *Статистическая*
- *Корреляционная*

# Прикладные цели исследования зависимостей

1. Установление самого факта наличия или отсутствия статистически значимой связи между У и Х
2. Прогноз неизвестных значений результирующих показателей по заданным значениям Х.
3. Выявление причинных связей между переменными Х и результирующими показателями У.

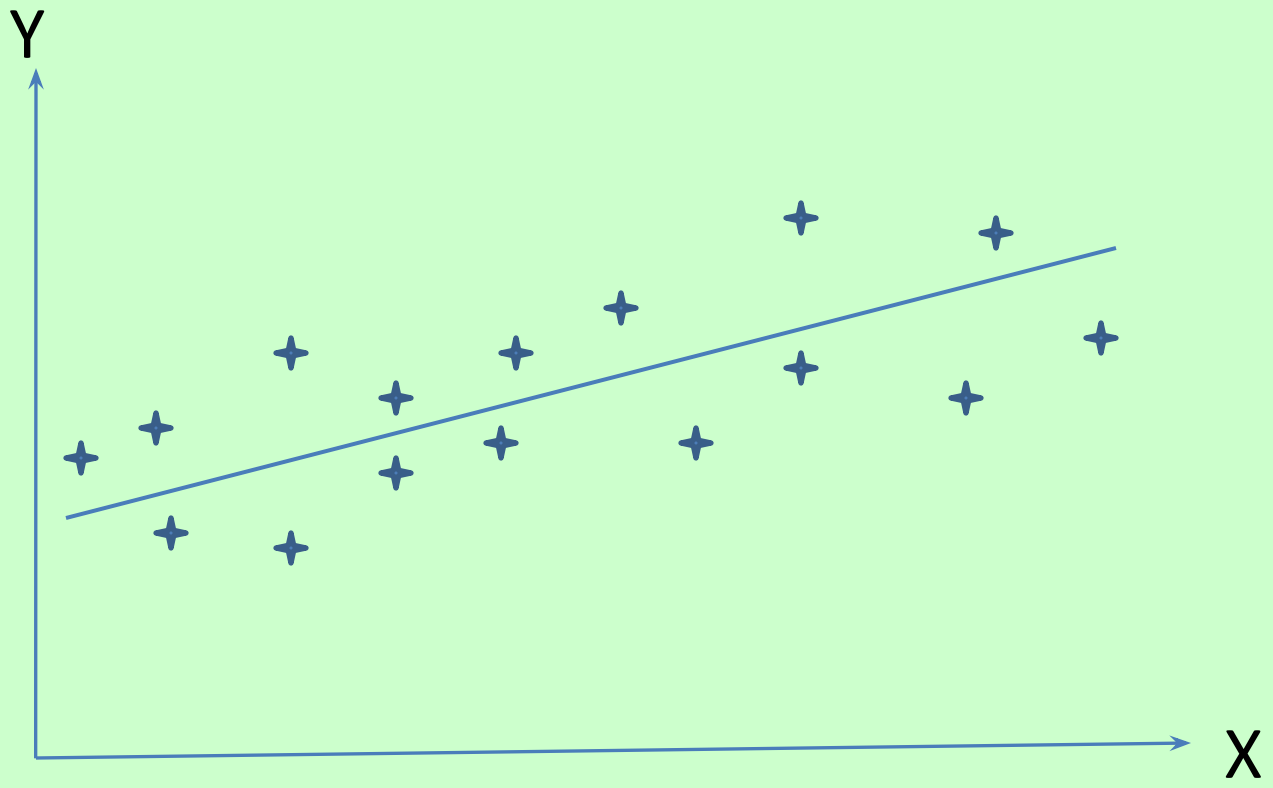
# Методы выявления наличия связи, ее характера и направления

- *приведения параллельных рядов данных*
- *аналитических группировок*
- *графический*
- *метод корреляции*

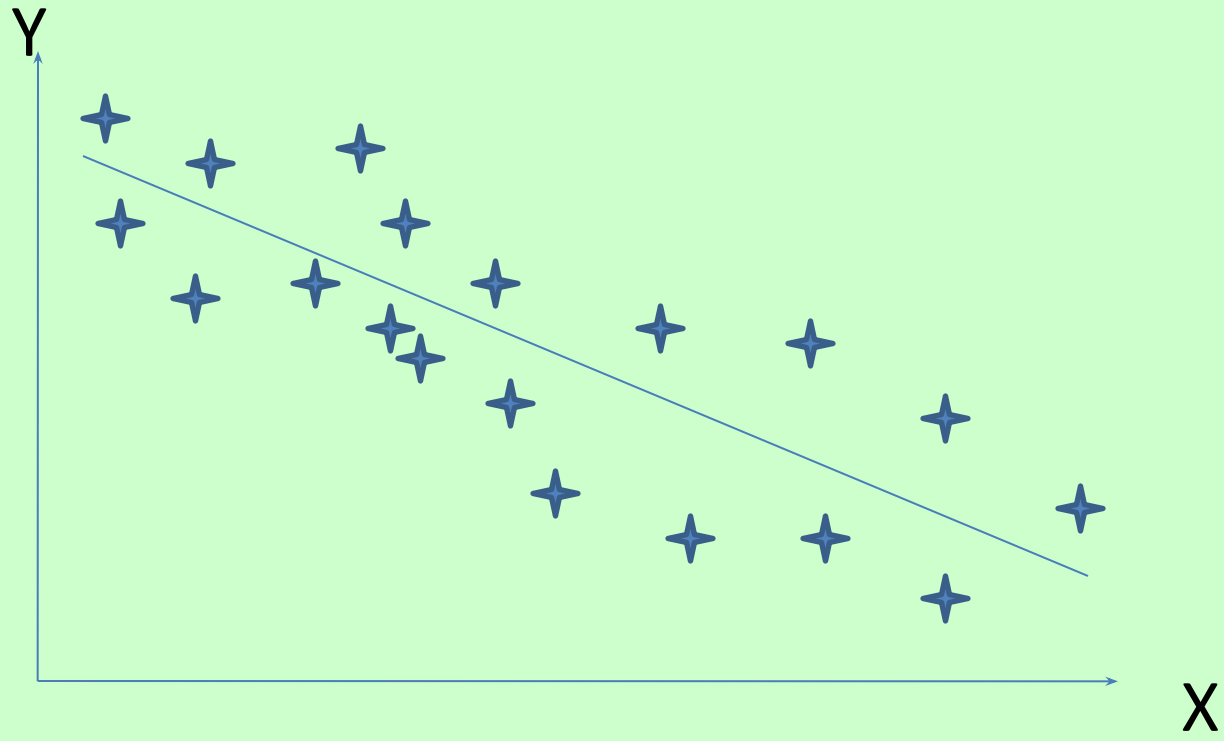
# Классификация связей

1. *по направлению связи*:
   - прямые
   - обратные
2. *по форме связи*:
   - линейные
   - нелинейные
3. *по количеству факторов*:
   - однофакторные
   - многофакторные

# Линейная корреляционная зависимость переменной Y от переменной X (положительная связь)



# Отрицательная линейная зависимость



# Линейный коэффициент корреляции

$$r_{xy} = \frac{\overline{xy} - \bar{x} \times \bar{y}}{\sigma_x \sigma_y}$$

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{(n\sum x_i^2 - (\sum x_i)^2)\cdot(n\sum y_i^2 - (\sum y_i)^2)}}$$

| Величина показателя связи | Характер связи |
| --- | --- |
| До ± 0,3 | Практически отсутствует |
| ±0,3 - ± 0,5 | Слабая |
| ±0,5 - ± 0,7 | Умеренная |
| ±0,7 - ± 1,0 | сильная |

# Непараметрические методы корреляционного анализа

## Коэффициенты, применяемые для характеристики тесноты связи между признаками разных типов

- **Ранговый коэффициент Спирмена кач/кол**
- **Ранговый коэффициент Кендела кач/кол**
- **Коэффициент взаимной сопряженности Пирсона, Чупрова кач**
- **Коэффициент ассоциации и контингенции кач**
- **Бисериальный коэффициент кач**
- **Коэффициент конкордации кач**

# Коэффициент корреляции рангов Спирмена

$$\rho = 1 - \frac{6 \cdot \sum d_i}{n(n^2 - 1)}$$

# Коэффициент корреляции рангов Кендела

$$\tau = \frac{2 \cdot S}{n(n-1)}$$

$$S = P - Q$$

# Коэффициенты взаимной сопряженности Пирсона, Чупрова

$$K_{П} = \sqrt{\frac{\varphi^2}{1+\varphi^2}} \qquad K_{Ч} = \sqrt{\frac{\varphi^2}{\sqrt{(K_1 - 1) * (K_2 - 1)}}}$$

$$\varphi^2 = \sum \frac{n_{xy}^2}{n_x n_y} - 1$$

$n_{xy}$ - *частота каждой клетки таблицы взаимной сопряженности*

$n_x$, $n_y$ *-итоговые частоты соответствующих строк и столбцов*

$K_1$, $K_2$ *–число строк и столбцов*

# Коэффициент ассоциации и контингенции

$$\kappa_{асс} = \frac{ad - bc}{ad + bc}$$

$$\kappa_{конт} = \frac{ad - bc}{\sqrt{(a+b)\cdot(b+d)\cdot(a+c)\cdot(c+d)}}$$

# Пример

| **Группы сотрудников** | **Средний балл по сравнению с предыдущей аттестацией** | | **Всего** |
|---|---|---|---|
| | Не изменился и возрос | снизился | |
| Прошедшие квалификацию | **163** (a) | **77** (в) | **240** (a+в) |
| Непрошедшие квалификацию | **46** (c) | **34** (d) | 80 (c+d) |
| **Всего:** | 209 | 111 | 320 |

# Точечный бисериальный коэффициент корреляции

$$r_{pb} = \frac{\bar{x}_1 - \bar{x}_0}{s_x}\sqrt{\frac{n_1 n_2}{n(n-1)}}$$

$\bar{x}_1$ – среднее значение по $X$ объектов со значением «единица» по $Y$;

$\bar{x}_0$ – среднее значение по $X$ объектов со значением «ноль» по $Y$;

$s_x$ – среднее квадратическое отклонение всех значений по $X$;

$n_1$ – число объектов «единица» по $Y$, $n_0$ – число объектов «ноль» по $Y$;

$n = n_1 + n_0$ – объем выборки.

# Рангово-бисериальный коэффициент корреляции

$$r_{rb} = \frac{2}{n}(\overline{X}_1 - \overline{X}_0)$$

Здесь $\overline{X}_1$ – средний ранг объектов, имеющих единицу по $Y$; $\overline{X}_0$ – средний ранг объектов с нулем по $Y$, $n$ – объем выборки.

# Коэффициент конкордации (согласованности) Кендалла

$$W = \frac{12 \cdot \sum_{i=1}^{n} D_i^2}{m^2 (n^3 - n)}$$

# Модель взаимосвязи показателей таможенной статистики

$$Y_i = \varphi(X_i) + \acute{\varepsilon}_i,$$

где $Y_i$ –значение результирующей переменной Y в i – том наблюдении;

$X_i$ – значение фактора X в i – том наблюдении;

$X = (X_1, X_2, \ldots, X_m)$-в общем случае вектор фактор;

m – количество компонентов вектора - фактора;

$\acute{\varepsilon}_i$ – значение случайной составляющей $\acute{\varepsilon}$ в i – том наблюдении (остатки);

i=1,2,…,n.

## Основные предпосылки применения регрессионного анализа:

- Достаточный объем наблюдений (не менее (8-10 единиц).
- Однородность изучаемых единиц.
- Случайная составляющая модели **ε** (остатки) имеет нормальное распределение с математическим ожиданием, равным нулю и постоянной дисперсией (Остатки **ε** не должны зависеть от значений фактора **X** .)
- Остатки $\varepsilon_i$ должны быть некоррелированы между собой.

# Формы регрессии

1. Регрессия парная.
2. Множественная регрессия.
3. Линейная регрессия.
4. Нелинейная регрессия относительно включенных в уравнение переменных, но линейная по параметрам.
5. Нелинейная регрессия, отличающаяся нелинейностью по оцениваемым параметрам.

# Этапы построения регрессионных моделей

1. **Выбор формулы связи переменных Y и X** :

**Y=φ (X)** (спецификация уравнения регрессии).

2. **Оценка параметров уравнения регрессии** и проверка надежности полученных оценок (параметризация уравнения регрессии).

3. **Статистический анализ модели: оценка точности и адекватности модели** (определение статистической значимости коэффициента детерминации, исследование случайной составляющей έ).

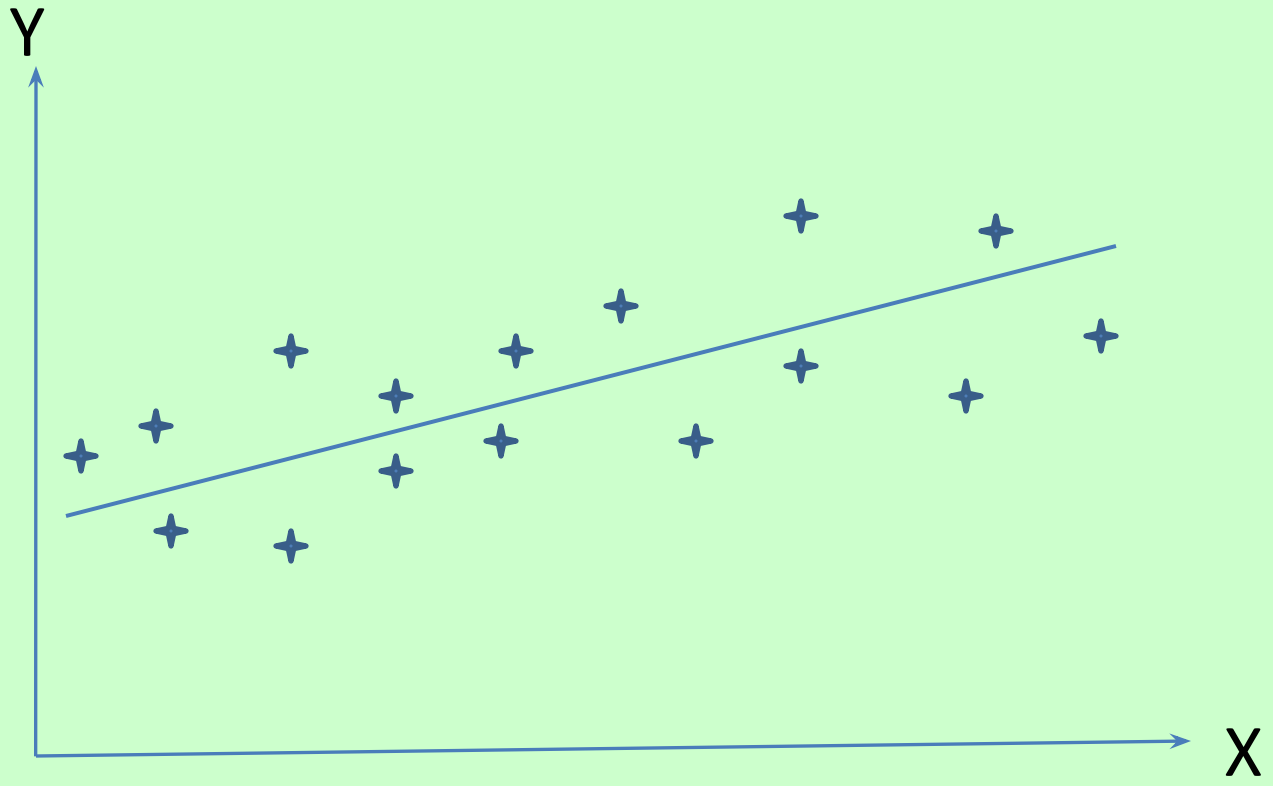
# Анализ взаимосвязи

- 1. Изобразить диаграмму, сформулировать гипотезу о форме связи.
- 2. Найти параметры уравнения линейной регрессии
- 3. Оценить статистическую значимость коэффициента регрессии, используя t-критерий Стьюдента
- 4. Рассчитать границы доверительного интервала для b
- 5. Вычислить коэффициенты корреляции, детерминации.
- 6. Выполнить прогноз

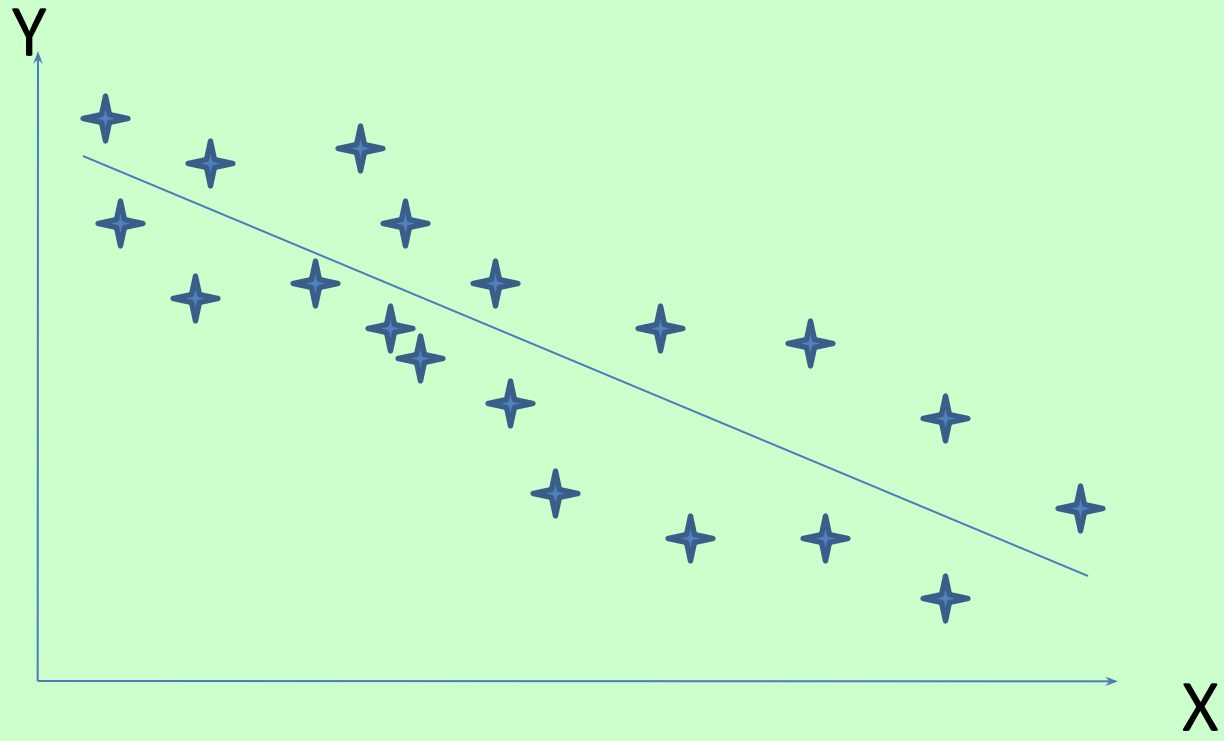
# 1.Графический анализ
# Линейная корреляционная зависимость

## переменной Y от переменной X

**(положительная связь)**



# Отрицательная линейная зависимость



**Связи нелинейного характера могут быть отображены функциями разного вида:**

$\hat{y} = ax^b$ - степенной ;

$\hat{y} = a + b \cdot \log x$ - логарифмической;

$\hat{y} = ab^x$ - показательной ;

$\hat{y} = a + \frac{b}{x}$ - гиперболической и др.

# 2. Линейное уравнение регрессии

$$y = a + bx$$

# Метод наименьших квадратов

$$S = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 \Rightarrow \min$$

$$\sum_{i=1}^{n} (y_i - (\tilde{a} + \tilde{b}\, x_i))^2 \Rightarrow \min$$

# Система нормальных уравнений

$$\sum_{i=1}^{n} y_i = \tilde{a} n + \tilde{b} \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} y_i x_i = \tilde{a} \sum_{i=1}^{n} x_i + \tilde{b} \sum_{i=1}^{n} x_i^2$$

# Расчетная таблица:

| | Месяц | Оборот млрд. долл. x | Таможенные платежи млрд. долл. y | xy | $x^2$ | $y^2$ | y(x) |
|---|---|---|---|---|---|---|---|
| 1 | январь | | | | | | |
| 2 | февраль | | | | | | |
| 3 | март | | | | | | |
| 4 | апрель | | | | | | |
| 5 | май | | | | | | |
| 6 | июнь | | | | | | |
| 7 | июль | | | | | | |
| 8 | август | | | | | | |
| 9 | сентябрь | | | | | | |
| 10 | октябрь | | | | | | |
| 11 | ноябрь | | | | | | |
| 12 | декабрь | | | | | | |

# Оценки параметров

$$b = \frac{\overline{yx} - \overline{y} \cdot \overline{x}}{\overline{x^2} - (\overline{x})^2}$$

- Коэффициент регресс

$$a = \overline{y} - b \cdot \overline{x}$$

# 4. Оценка статистической значимости коэффициента регрессии

Выдвигаем гипотезу Ho:b=0 об отсутствии влияния фактора на отклик

1) Стандартная ошибка

$$SE_b = \sqrt{\frac{\sum e^2}{(N-1)(N-2)\cdot \delta^2{}_x}}$$

$$\delta_x = \sqrt{\frac{N}{(N-1)}\cdot(\overline{x^2} - \overline{x}^2)}$$

N – число наблюдений

# Расчетная таблица:

| | Месяц | Оборот млрд. долл. x | Таможенные платежи млрд. долл. y | xy | $x^2$ | $y^2$ | y(x) | e=y-y(x) | $e^2$ | \|e/y\| *100 % |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | январь | | | | | | | | | |
| 2 | февраль | | | | | | | | | |
| 3 | март | | | | | | | | | |
| 4 | апрель | | | | | | | | | |
| 5 | май | | | | | | | | | |
| 6 | июнь | | | | | | | | | |
| 7 | июль | | | | | | | | | |
| 8 | август | | | | | | | | | |
| 9 | сентябрь | | | | | | | | | |
| 10 | октябрь | | | | | | | | | |
| 11 | ноябрь | | | | | | | | | |
| 12 | декабрь | | | | | | | | | |
| сумма | | | | | | | | | | |
| среднее | | | | | | | | | | |

- Рассчитываем фактическое значение t-критерия Стьюдента и сравниваем с табличным значением на уровне значимости α=0,05 и числа степеней свободы N-2=12-2=10

$$t_b = \frac{|b|}{SE_b}$$

$t_b > t_{табл}$ – гипотеза Но отклоняется

# Критические значения критерия t-Стьюдента

| df | α 0,10 | 0,05 | 0,01 | 0,001 | df | α 0,10 | 0,05 | 0,01 | 0,001 | df | α 0,10 | 0,05 | 0,01 | 0,001 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 6,314 | 12,70 | 63,65 | 636,61 | **31** | 1,696 | 2,040 | 2,744 | 3,633 | **61** | 1,670 | 2,000 | 2,659 | 3,457 |
| **2** | 2,920 | 4,303 | 9,925 | 31,602 | **32** | 1,694 | 2,037 | 2,738 | 3,622 | **62** | 1,670 | 1,999 | 2,657 | 3,454 |
| **3** | 2,353 | 3,182 | 5,841 | 12,923 | **33** | 1,692 | 2,035 | 2,733 | 3,611 | **63** | 1,669 | 1,998 | 2,656 | 3,452 |
| **4** | 2,132 | 2,776 | 4,604 | 8,610 | **34** | 1,691 | 2,032 | 2,728 | 3,601 | **64** | 1,669 | 1,998 | 2,655 | 3,449 |
| **5** | 2,015 | 2,571 | 4,032 | 6,869 | **35** | 1,690 | 2,030 | 2,724 | 3,591 | **65** | 1,669 | 1,997 | 2,654 | 3,447 |
| **6** | 1,943 | 2,447 | 3,707 | 5,959 | **36** | 1,688 | 2,028 | 2,719 | 3,582 | **66** | 1,668 | 1,997 | 2,652 | 3,444 |
| **7** | 1,895 | 2,365 | 3,499 | 5,408 | **37** | 1,687 | 2,026 | 2,715 | 3,574 | **67** | 1,668 | 1,996 | 2,651 | 3,442 |
| **8** | 1,860 | 2,306 | 3,355 | 5,041 | **38** | 1,686 | 2,024 | 2,712 | 3,566 | **68** | 1,668 | 1,995 | 2,650 | 3,439 |
| **9** | 1,833 | 2,262 | 3,250 | 4,781 | **39** | 1,685 | 2,023 | 2,708 | 3,558 | **69** | 1,667 | 1,995 | 2,649 | 3,437 |
| **10** | 1,812 | **2,228** | 3,169 | 4,587 | **40** | 1,684 | 2,021 | 2,704 | 3,551 | **70** | 1,667 | 1,994 | 2,648 | 3,435 |
| **11** | 1,796 | 2,201 | 3,106 | 4,437 | **41** | 1,683 | 2,020 | 2,701 | 3,544 | **71** | 1,667 | 1,994 | 2,647 | 3,433 |
| **12** | 1,782 | 2,179 | 3,055 | 4,318 | **42** | 1,682 | 2,018 | 2,698 | 3,538 | **72** | 1,666 | 1,993 | 2,646 | 3,431 |
| **13** | 1,771 | 2,160 | 3,012 | 4,221 | **43** | 1,681 | 2,017 | 2,695 | 3,532 | **73** | 1,666 | 1,993 | 2,645 | 3,429 |
| **14** | 1,761 | 2,145 | 2,977 | 4,140 | **44** | 1,680 | 2,015 | 2,692 | 3,526 | **74** | 1,666 | 1,993 | 2,644 | 3,427 |
| **15** | 1,753 | 2,131 | 2,947 | 4,073 | **45** | 1,679 | 2,014 | 2,690 | 3,520 | **75** | 1,665 | 1,992 | 2,643 | 3,425 |
| **16** | 1,746 | 2,120 | 2,921 | 4,015 | **46** | 1,679 | 2,013 | 2,687 | 3,515 | **76** | 1,665 | 1,992 | 2,642 | 3,423 |
| **17** | 1,740 | 2,110 | 2,898 | 3,965 | **47** | 1,678 | 2,012 | 2,685 | 3,510 | **78** | 1,665 | 1,991 | 2,640 | 3,420 |
| **18** | 1,734 | 2,101 | 2,878 | 3,922 | **48** | 1,677 | 2,011 | 2,682 | 3,505 | **79** | 1,664 | 1,990 | 2,639 | 3,418 |
| **19** | 1,729 | 2,093 | 2,861 | 3,883 | **49** | 1,677 | 2,010 | 2,680 | 3,500 | **80** | 1,664 | 1,990 | 2,639 | 3,416 |
| **20** | 1,725 | 2,086 | 2,845 | 3,850 | **50** | 1,676 | 2,009 | 2,678 | 3,496 | **90** | 1,662 | 1,987 | 2,632 | 3,402 |
| **21** | 1,721 | 2,080 | 2,831 | 3,819 | **51** | 1,675 | 2,008 | 2,676 | 3,492 | **100** | 1,660 | 1,984 | 2,626 | 3,390 |
| **22** | 1,717 | 2,074 | 2,819 | 3,792 | **52** | 1,675 | 2,007 | 2,674 | 3,488 | **110** | 1,659 | 1,982 | 2,621 | 3,381 |
| **23** | 1,714 | 2,069 | 2,807 | 3,768 | **53** | 1,674 | 2,006 | 2,672 | 3,484 | **120** | 1,658 | 1,980 | 2,617 | 3,373 |
| **24** | 1,711 | 2,064 | 2,797 | 3,745 | **54** | 1,674 | 2,005 | 2,670 | 3,480 | **130** | 1,657 | 1,978 | 2,614 | 3,367 |
| **25** | 1,708 | 2,060 | 2,787 | 3,725 | **55** | 1,673 | 2,004 | 2,668 | 3,476 | **140** | 1,656 | 1,977 | 2,611 | 3,361 |
| **26** | 1,706 | 2,056 | 2,779 | 3,707 | **56** | 1,673 | 2,003 | 2,667 | 3,473 | **150** | 1,655 | 1,976 | 2,609 | 3,357 |

# 5. Рассчитываем Границы 95-процентного доверительного интервала для коэффициента регрессии

$$Н.гр. = b - t_{табл} * SE_b$$

$$В.гр. = b + t_{табл} * SE_b$$

# 6. Рассчитываем Коэффициент корреляции

$$r = b\frac{\delta_x}{\delta_y}$$

$$\delta_x = \sqrt{\frac{N}{(N-1)} \cdot (\overline{x^2} - \bar{x}^2)}$$

$$\delta_y = \sqrt{\frac{N}{(N-1)} \cdot (\overline{y^2} - \bar{y}^2)}$$

# Степень тесноты связи

| **Величина показателя связи** | **Характер связи** |
|---|---|
| До ± 0,3 | Практически отсутствует |
| ±0,3 - ± 0,5 | Слабая |
| ±0,5 - ± 0,7 | Умеренная |
| ±0,7 - ± 1,0 | сильная |

# 7. Оценка адекватности уравнения регрессии

- Теоретический коэффициент детерминации

$$R_{yx}^2 = r^2$$

$R^2$>30% - прогнозировать по модели целесообразно

# 8. Оценка значимости уравнения регрессии

Выдвигаем гипотезу Ho:b=0 о статистической незначимости уравнения регрессии и коэффициента детерминации

Рассчитываем фактическое значение F-критерия Фишера и сравниваем с табличным значением на уровне значимости α=0,05 и числе степеней свободы 1 и N-2=12-2=10

$$F_{факт} = \frac{R^2}{1 - R^2} \cdot (N - 2)$$

$F_{факт} > F_{табл}$ – гипотеза отклоняется

# Критические значения критерия F-Фишера

α=0,05

| | Степени свободы для числителя | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **10** | **12** | **24** | **?** |
| **3** | 10,128 | 9,552 | 9,277 | 9,117 | 9,013 | 8,941 | 8,887 | 8,845 | 8,785 | 8,745 | 8,638 | 8,527 |
| **5** | 6,608 | 5,786 | 5,409 | 5,192 | 5,050 | 4,950 | 4,876 | 4,818 | 4,735 | 4,678 | 4,527 | 4,366 |
| **7** | 5,591 | 4,737 | 4,347 | 4,120 | 3,972 | 3,866 | 3,787 | 3,726 | 3,637 | 3,575 | 3,410 | 3,231 |
| **10** | **4,965** | 4,103 | 3,708 | 3,478 | 3,326 | 3,217 | 3,135 | 3,072 | 2,978 | 2,913 | 2,737 | 2,539 |
| **11** | 4,844 | 3,982 | 3,587 | 3,357 | 3,204 | 3,095 | 3,012 | 2,948 | 2,854 | 2,788 | 2,609 | 2,406 |
| **12** | 4,747 | 3,885 | 3,490 | 3,259 | 3,106 | 2,996 | 2,913 | 2,849 | 2,753 | 2,687 | 2,505 | 2,297 |
| **13** | 4,667 | 3,806 | 3,411 | 3,179 | 3,025 | 2,915 | 2,832 | 2,767 | 2,671 | 2,604 | 2,420 | 2,208 |
| **14** | 4,600 | 3,739 | 3,344 | 3,112 | 2,958 | 2,848 | 2,764 | 2,699 | 2,602 | 2,534 | 2,349 | 2,132 |
| **15** | 4,543 | 3,682 | 3,287 | 3,056 | 2,901 | 2,790 | 2,707 | 2,641 | 2,544 | 2,475 | 2,288 | 2,067 |
| **16** | 4,494 | 3,634 | 3,239 | 3,007 | 2,852 | 2,741 | 2,657 | 2,591 | 2,494 | 2,425 | 2,235 | 2,011 |
| **18** | 4,414 | 3,555 | 3,160 | 2,928 | 2,773 | 2,661 | 2,577 | 2,510 | 2,412 | 2,342 | 2,150 | 1,918 |
| **20** | 4,351 | 3,493 | 3,098 | 2,866 | 2,711 | 2,599 | 2,514 | 2,447 | 2,348 | 2,278 | 2,082 | 1,844 |
| **30** | 4,171 | 3,316 | 2,922 | 2,690 | 2,534 | 2,421 | 2,334 | 2,266 | 2,165 | 2,092 | 1,887 | 1,624 |
| **40** | 4,085 | 3,232 | 2,839 | 2,606 | 2,449 | 2,336 | 2,249 | 2,180 | 2,077 | 2,003 | 1,793 | 1,511 |
| **50** | 4,034 | 3,183 | 2,790 | 2,557 | 2,400 | 2,286 | 2,199 | 2,130 | 2,026 | 1,952 | 1,737 | 1,440 |
| **70** | 3,978 | 3,128 | 2,736 | 2,503 | 2,346 | 2,231 | 2,143 | 2,074 | 1,969 | 1,893 | 1,674 | 1,355 |
| **100** | 3,936 | 3,087 | 2,696 | 2,463 | 2,305 | 2,191 | 2,103 | 2,032 | 1,927 | 1,850 | 1,627 | 1,286 |
| **200** | 3,888 | 3,041 | 2,650 | 2,417 | 2,259 | 2,144 | 2,056 | 1,985 | 1,878 | 1,801 | 1,572 | 1,192 |
| **oo** | 3,843 | 2,998 | 2,607 | 2,374 | 2,216 | 2,100 | 2,011 | 1,940 | 1,833 | 1,754 | 1,519 | |

# 9. Прогноз ожидаемого значения y по уравнению регрессии

- Средняя абсолютная ошибка прогноза

$$\mathbf{MAPE = (|e/y|*100)/N}$$

# 9. Прогноз ожидаемого значения y по уравнению регрессии

- Точечный

$$y_f = a + bx_f$$

- Интервальный

$$\text{Н.гр.} = y_f - t_{\text{табл}} * SE_f$$

$$\text{В.гр.} = y_f + t_{\text{табл}} * SE_f$$

# Стандартная ошибка прогноза

$$SE_f = \sqrt{\frac{\sum e^2}{N-2}\left(1+\frac{1}{N}+\frac{(x_f-\bar{x})^2}{(N-1)\cdot\delta^2{}_x}\right)}$$