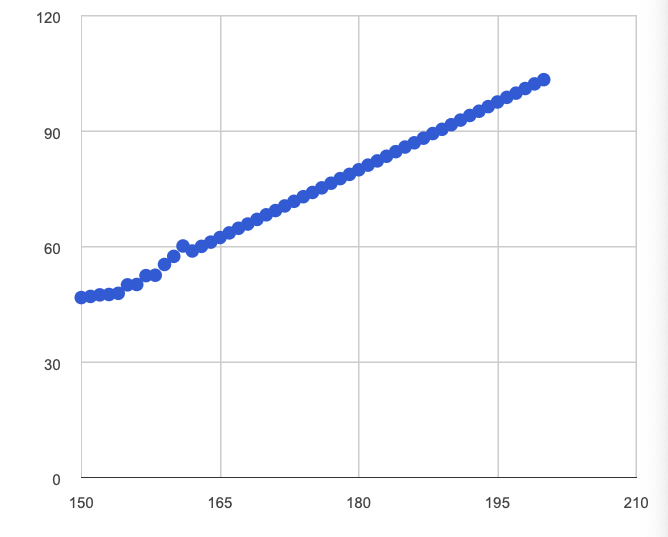
1. Описание данных

Дано : рост и вес людей

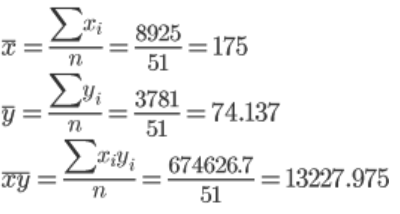
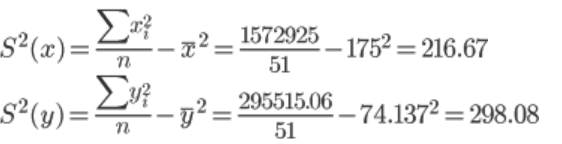
**Задача работы:** Построение модели регрессии для данной выборки. Оценка качества уравнения регрессии.

1. Корреляционное поле

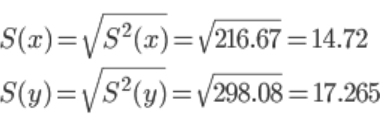


Совокупность точек результативного и факторного признаков называется **полем корреляции**

**Выборочные средние Выборочные дисперсии**

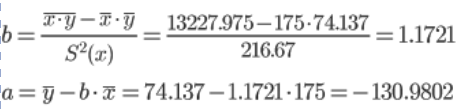
 

**Среднеквадрат.отклонение:**



Стандартная ошибка регрессии рассматривается в качестве меры разброса данных наблюдений от смоделированных значений. Чем меньше значение стандартной ошибки регрессии, тем качество модели выше.   
Sa - стандартное отклонение случайной величины a.

Коэффициент корреляции b можно находить по формуле, не решая систему непосредственно:

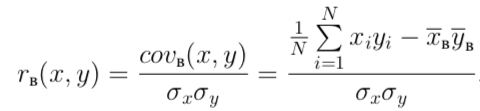
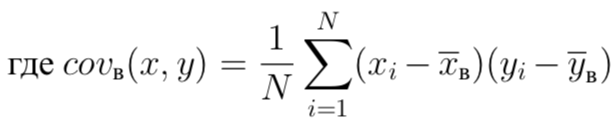


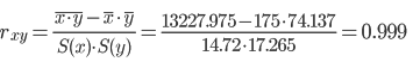
*Ковариация*



1. Выборочный линейный коэффициент корреляции (показатель тесноты связи)

Выборочный коэффициент корреляции случайных величин X и Y вычисляется по формуле :

**** ****

****

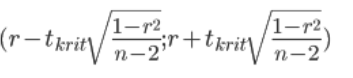
Линейный коэффициент корреляции принимает значения от –1 до +1

Связи между признаками могут быть слабыми и сильными (тесными). Их критерии оцениваются по шкале Чеддока:

0.1 < rxy < 0.3: слабая;    
0.3 < rxy < 0.5: умеренная;    
0.5 < rxy < 0.7: заметная;    
0.7 < rxy < 0.9: высокая;    
0.9 < rxy < 1: весьма высокая;

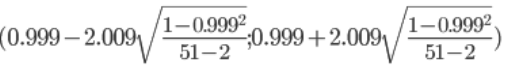
В нашем примере связь между признаком Y и фактором X весьма высокая и прямая. 

* 1. Доверительный интервал для коэффициента корреляции  
     (Интервальная оценка для коэффициента корреляции)

****

Для оценки значимости коэффициента корреляции используют критерий Стьюдента. [По таблице Стьюдента](https://math.semestr.ru/corel/table-student.php) находим:

tкрит(n-m-1;α/2) = tкрит(49;0.025) = 2.009

****

r с (0.989 ; 1)

Проверка значимости

Выборочный коэффициент корреляции является статистической оценкой коэффициента корреляции генеральной совокупности. Если случайные величины зависимы линейно, то |r(x, y)| = 1

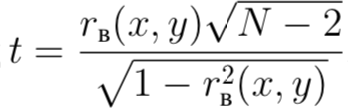
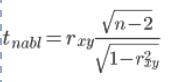
Если случайные величины независимы, то коэффициент корреляции равен нулю, однако обратное не верно. Между величинами может существовать зависимость, которая не выражается линейно. В этом случае проверяется гипотеза о значимости коэффициента

Она имеет распределение, близкое к распределению Стьюдента с N − 2 степенями свободы

Выдвигаем гипотезы:

H0: rxy = 0, нет линейной взаимосвязи между переменными;   
H1: rxy ≠ 0, есть линейная взаимосвязь между переменными;

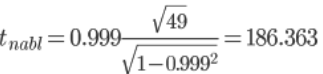
Для того чтобы при уровне значимости α проверить нулевую гипотезу (H : ρ(x, y) = 0, где ρ - коэффициент корреляции генеральной совокупности) о равенстве нулю генерального коэффициента корреляции нормальной двумерной случайной величины при конкурирующей гипотезе H1 ≠ 0, надо вычислить наблюдаемое значение критерия (величина случайной ошибки)



критическую точку tкрит двусторонней критической области.

Если tнабл < tкрит нет оснований отвергнуть нулевую гипотезу.

Если |tнабл| > tкрит — нулевую гипотезу отвергают. 



Если |tнабл| > tкритич, то полученное значение коэффициента корреляции признается значимым (нулевая гипотеза, утверждающая равенство нулю коэффициента корреляции, отвергается).

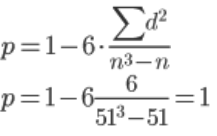
=> отклоняем гипотезу о равенстве 0 коэффициента корреляции

Kоэффициент корреляции статистически - значим

1. Ранговые коэффициенты корреляции

** **

По формуле вычислим коэффициент ранговой корреляции Спирмена:

****

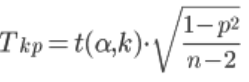
Связь между признаком Y и фактором X сильная и прямая

* 1. Оценка коэффициента ранговой корреляции Спирмена

Значимость коэффициента ранговой корреляции Спирмена

Для того чтобы при уровне значимости α проверить нулевую гипотезу о равенстве нулю генерального коэффициента ранговой корреляции Спирмена при конкурирующей гипотезе Hi. p ≠ 0

надо вычислить критическую точку:

****

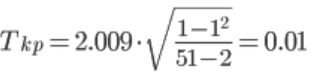
где n - объем выборки , p - выборочный коэффициент ранговой корреляции Спирмена:

t(α, к) - критическая точка двусторонней критической области, которую находят по таблице критических точек распределения Стьюдента, по уровню значимости α и числу степеней свободы k = n-2.

Если |p| < Тkp - нет оснований отвергнуть нулевую гипотезу.

Ранговая корреляционная связь между качественными признаками не значима. Если |p| > Tkp - нулевую гипотезу отвергают. Между качественными признаками существует значимая ранговая корреляционная связь.

По таблице Стьюдента находим t(α/2, k) = (0.05/2;49) = 2.009

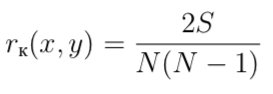
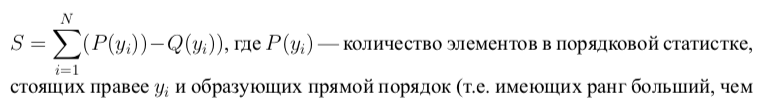
****

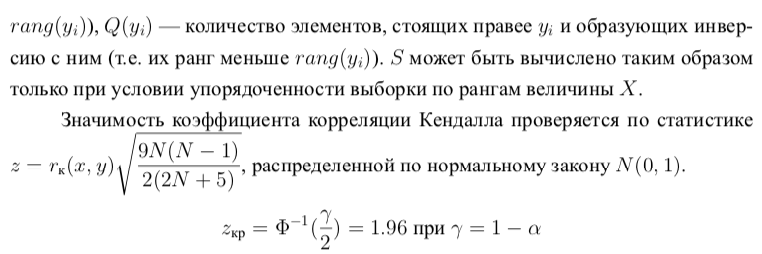
Поскольку Tkp < p, то отклоняем гипотезу о равенстве 0 коэффициента ранговой корреляции Спирмена.

Другими словами, коэффициент ранговой корреляции статистически - значим и ранговая корреляционная связь между оценками по двум тестам значимая.

1. Коэффициент Кендэла

Ранговый коэффициент корреляции Кендалла вычисляется по формуле :

 (1)





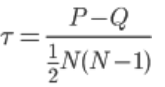
Далее S считаешь и подставляешь в (1)

Считаешь z

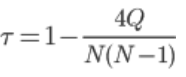
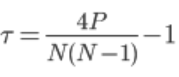
Если Z > 1,96 , то 

Др вариант

Присвоим ранги признаку Y и фактору X.   
Расположим объекты так, чтобы их ранги по X представили натуральный ряд. Так как оценки, приписываемые каждой паре этого ряда, положительные, значения «+1», входящие в Р, будут порождаться только теми парами, ранги которых по Y образуют прямой порядок.   
Их легко подсчитать, сопоставляя последовательно ранги каждого объекта в ряду Y с стальными.



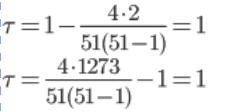
В общем случае расчет τ (точнее Р или Q) даже для N порядка 10 оказывается громоздким. Покажем, как упростить вычисления.

 = 

Упорядочим данные по X  
В ряду Y справа от 1 расположено 50 рангов, превосходящих 1, следовательно, 1 породит в Р слагаемое 50.   
Справа от 2 стоят 49 ранга, превосходящих 2 (это 3, 4, 5, 6, 7, 8, 9, 10, 11, 14, 12, 13, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51), т.е. в Р войдет 49 и т.д. В итоге Р = 1273 и с использованием формул имеем:

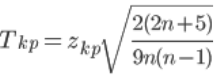


По упрощенным формулам:



Для того чтобы при уровне значимости α проверить нулевую гипотезу о равенстве нулю генерального коэффициента ранговой корреляции Кендалла при конкурирующей гипотезе Н1: τ ≠ 0 ,

надо вычислить критическую точку:



где n - объем выборки;

zkp - критическая точка двусторонней критической области, которую находят по таблице функции Лапласа по равенству :

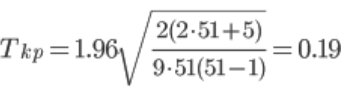
Если |τ| < Tkp — нет оснований отвергнуть нулевую гипотезу.

Ранговая корреляционная связь между качественными признаками незначима.

Если |τ| > Tkp — нулевую гипотезу отвергают. Между качественными признаками существует значимая ранговая корреляционная связь.

Найдем критическую точку zkp

По таблице Лапласа находим zkp = 1.96   
Найдем критическую точку:



Так как τ > Tkp — отвергаем нулевую гипотезу;

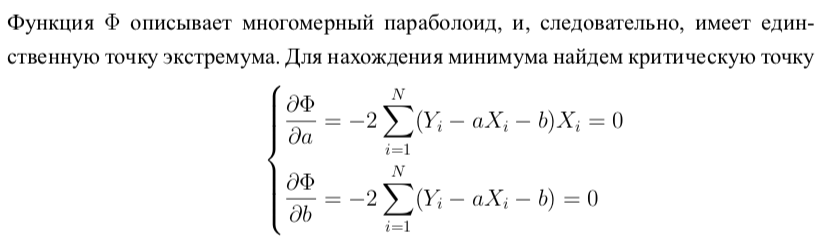
ранговая корреляционная связь между оценками по двум тестам значимая

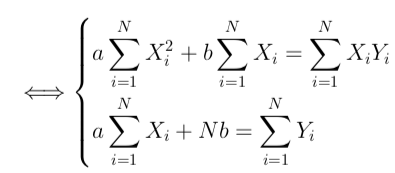
1. Линейное уравнение регрессии

На основании поля корреляции можно выдвинуть гипотезу (для генеральной совокупности) о том, что связь между всеми возможными значениями X и Y носит линейный характер.

Регрессией называется функция, определенная как условное математическое ожидание Y при условии X = x, т.е. f(x) = M(Y | X = x). Функция f(x) может иметь произвольный вид, нахождение этой функции в явном виде составляет суть регрессионного анализа.

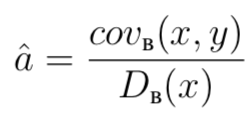
Уравнение линейной регрессии в общем виде записывается как y = ax + b. Оценку коэффициентов линейной регрессии производим по методу наименьших квадратов. Суть метода наименьших квадратов заключается в минимизации случайных отклонений от выбранного уравнения регрессии. Для этого рассмотрим многомерную функцию

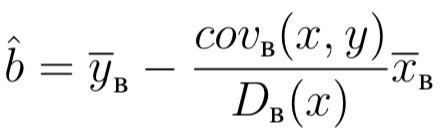




Выразив переменные(ниже) a и b и подставив известные выражения для выборочных

среднего, дисперсии и ковариации, получим выражение для оценок коэфиициентов уравнения линейной регрессии





**Линейное уравнение регрессии имеет вид y = bx + a + ε**

Система нормальных уравнений:

a·n + b·∑x = ∑y   
a·∑x + b·∑x2 = ∑y·x

Для наших данных система уравнений имеет вид:

51a + 8925·b = 3781  a = -130.98 **Уравнение регрессии:**8925·a + 1572925·b = 674626.7  b = 1.172 y = 1.1721 x -130.9802

Kоэффициенты регрессии *a* и *b* являются лишь оценками теоретических коэффициентов βi, а само уравнение отражает лишь общую тенденцию в поведении рассматриваемых переменных.

Коэффициент регрессии b = 1.172 показывает среднее изменение результативного показателя (в единицах измерения у) с повышением или понижением величины фактора х на единицу его измерения.

В данном примере с увеличением на 1 единицу y повышается в среднем на 1.172.

Коэффициент a = -130.98 формально показывает прогнозируемый уровень у, но только в том случае, если х=0 находится близко с выборочными значениями.

Но если х=0 находится далеко от выборочных значений х, то буквальная интерпретация может привести к неверным результатам, и даже если линия регрессии довольно точно описывает значения наблюдаемой выборки, нет гарантий, что также будет при экстраполяции влево или вправо.   
Подставив в уравнение регрессии соответствующие значения х, можно определить выровненные (предсказанные) значения результативного показателя y(x) для каждого наблюдения.

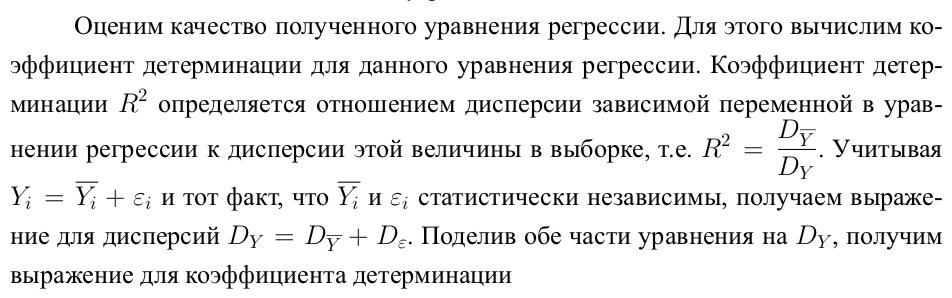
Связь между у и х определяет знак коэффициента регрессии b (если > 0 – прямая связь, иначе - обратная).

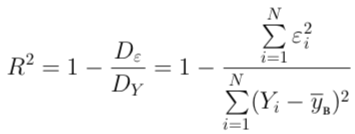
В нашем примере связь прямая.

21,4

* 1. Коэффициент детерминации



****

Квадрат (множественного) коэффициента корреляции называется коэффициентом детерминации, который показывает долю вариации результативного признака, объясненную вариацией факторного признака.

= 0,998

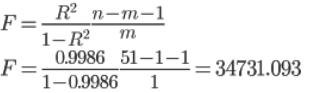
т.е. в 99.86% случаев изменения х приводят к изменению y. Другими словами - точность подбора уравнения регрессии - высокая. Остальные 0.14% изменения Y объясняются факторами, не учтенными в модели (а также ошибками спецификации)

* 1. F-статистика. Критерий Фишера

Оценка статистической значимости парной линейной регрессии производится по следующему алгоритму:

1. Выдвигается нулевая гипотеза о том, что уравнение в целом статистически незначимо: H0: R2=0 на уровне значимости α.

2. Далее определяют фактическое значение F-критерия:



3. Табличное значение определяется по таблицам распределения Фишера для заданного уровня значимости, принимая во внимание, что число степеней свободы для общей суммы квадратов (большей дисперсии) равно 1 и число степеней свободы остаточной суммы квадратов (меньшей дисперсии) при линейной регрессии равно n-2.   
Fтабл - 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.

4. Если фактическое значение F-критерия меньше табличного, то говорят, что нет основания отклонять нулевую гипотезу.   
В противном случае, нулевая гипотеза отклоняется и с вероятностью (1-α) принимается альтернативная гипотеза о статистической значимости уравнения в целом.

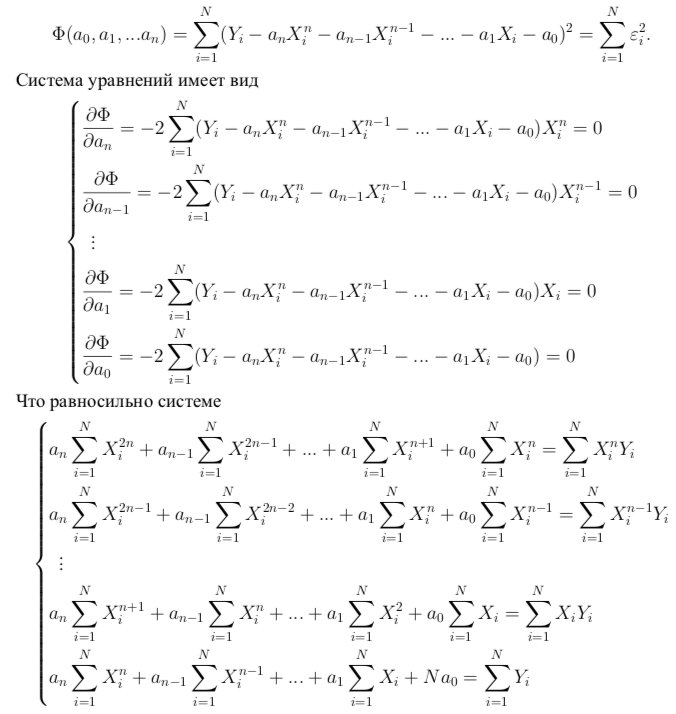
Табличное значение критерия со степенями свободы k1=1 и k2=49, Fтабл = 4

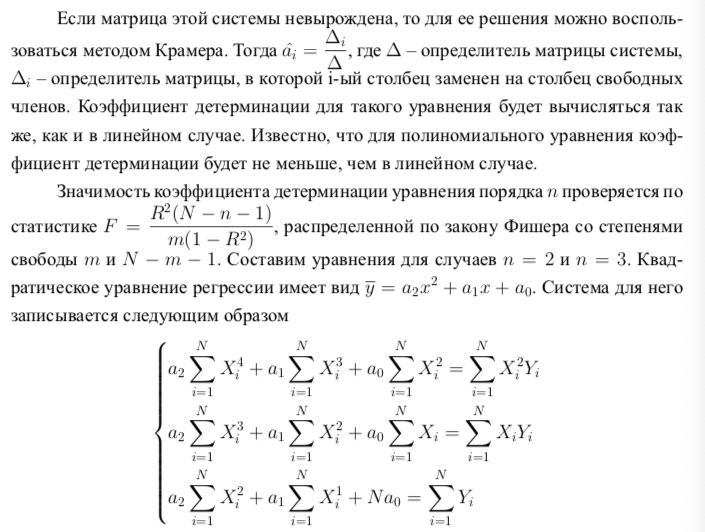
Поскольку фактическое значение F > Fтабл, то коэффициент детерминации статистически значим (найденная оценка уравнения регрессии статистически надежна).

1. Полиномиальная регрессия

Уравнение полиномиальной регрессии порядка n записывается в виде y = anxn + an−1xn−1 + ... + a1x + a0.

Для оценки коэффициентов уравнения полиномиальной регрессии также можно воспользоваться методом наименьших квадратов. Функция Φ в данном случае записывается как

****

****

|  |  |  |  |
| --- | --- | --- | --- |
| матрица для квадратного уравнения | | |  |
| 49867166665 | 279129375 | 1572925 | 121146300,3 |
| 279129375 | 1572925 | 8925 | 674631,4773 |
| 1572925 | 8925 | 51 | 3781,031818 |
|  |  |  |  |
| коэффициенты |  |  |  |
| a | 0,000083 |  |  |
| b | 1,14 |  |  |
| c | -128,44 |  |  |
|  |  |  |  |
| E^2 | 35,1551416 |  |  |
|  |  |  |  |
| R^2 | 0,99773254 |  |  |
|  |  |  |  |
| F | 21561,1141 |  |  |
|  |  |  |  |

****

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| матрица для кубичесткого уравнения | | |  |  |
| 1,6223E+15 | 8,96665E+12 | 49867166665 | 279129375 | 21889047091 |
| 8,96665E+12 | 49867166665 | 279129375 | 1572925 | 121146300,3 |
| 49867166665 | 279129375 | 1572925 | 8925 | 674631,4773 |
| 279129375 | 1572925 | 8925 | 51 | 3781,031818 |
|  |  |  |  |  |
| коэффициенты |  |  |  |  |
| a | 5,31506E-07 |  |  |  |
| b | 0,000005 |  |  |  |
| c | 1,17 |  |  |  |
| d | -131,101655 |  |  |  |
|  |  |  |  |  |
| E^2 | 385,5456666 |  |  |  |
|  |  |  |  |  |
| R^2 | 0,975132851 |  |  |  |
|  |  |  |  |  |
| F | 1921,471169 |  |  |  |

1. Показательная и степенная регрессии

На основании поля корреляции можно выдвинуть гипотезу (для генеральной совокупности) о том, что связь между всеми возможными значениями X и Y носит показательный характер.

Показательное уравнение регрессии имеет вид y = a\*bx

Оценочное уравнение регрессии (построенное по выборочным данным) будет иметь вид :

**y = a\*bx+ ε**

где ei – наблюдаемые значения ошибок εi, a и b соответственно оценки параметров α и β регрессионной модели

Здесь ε - случайная ошибка

Так как отклонения εi для каждого конкретного наблюдения i – случайны и их значения в выборке неизвестны, то:   
1) по наблюдениям xi и yi можно получить только оценки параметров α и β   
2) Оценками параметров α и β регрессионной модели являются соответственно величины а и b, которые носят случайный характер, т.к. соответствуют случайной выборке;

После линеаризации получим: **ln(y) = ln(a) + x ln(b)**

Для оценки параметров α и β - используют МНК (метод наименьших квадратов).   
МНК дает наилучшие оценки параметров уравнения регрессии. Но только в том случае, если выполняются определенные предпосылки относительно случайного члена (ε) и независимой переменной (x).

Система нормальных уравнений:

a·n + b·∑x = ∑y  ∑x = 8925 ∑x2 = 1572925  
a·∑x + b·∑x2 = ∑y·x  ∑y = ∑ln(y) = 218.1455 ∑y·x  = ln(y)\*x = 38356.237

Для наших данных система уравнений имеет вид:

51a + 8925·b = 218.145  b = 0.01636  
8925·a + 1572925·b = 38356.237  a = 1.4143

Уравнение регрессии :

**y = e1.4143407250193\*e0.01636x = 4.11377 \* 1.01649x**

Коэффициенты регрессии *a* и *b* являются лишь оценками теоретических коэффициентов βi, а само уравнение отражает лишь общую тенденцию в поведении рассматриваемых переменных.

На основании поля корреляции можно выдвинуть гипотезу (для генеральной совокупности) о том, что связь между всеми возможными значениями X и Y носит степенной характер.

0,03034328

Степенное уравнение регрессии имеет вид y = a\*xb

После линеаризации получим: ln(y) = ln(a) + b ln(x)

Для оценки параметров α и β - используют МНК (метод наименьших квадратов).   
МНК дает наилучшие оценки параметров уравнения регрессии. Но только в том случае, если выполняются определенные предпосылки относительно случайного члена (ε) и независимой переменной (x)

Система нормальных уравнений:

a·n + b·∑x = ∑·y  ∑ ln(x)i = 114.3161 ∑ ln(y)i = 94.7394  
a·∑x + b·∑x2 = ∑y·x  ∑ ln(x)2i =256.307 ∑ ln(x)i • ln(y)I = 212.554

Для наших данных система уравнений имеет вид :

51a + 114.316·b = 94.739  b = 2.8582

114.316·a + 256.307·b = 212.554  a = -4.5489

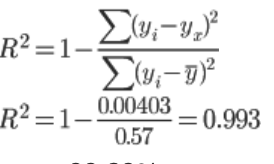
**Уравнение регрессии (эмпирическое уравнение регрессии):**

y = 10-4.5489089242114x2.8582 = 2.8E-5 \* x2.8582

Эмпирические коэффициенты регрессии *a* и *b* являются лишь оценками теоретических коэффициентов βi, а само уравнение отражает лишь общую тенденцию в поведении рассматриваемых переменных.

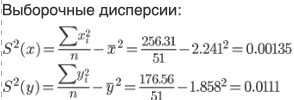
**Индекс детерминации**

Величину R2 (равную отношению объясненной уравнением регрессии дисперсии результата *у* к общей дисперсии *у*) для нелинейных связей называют **индексом детерминации**.



т.е. в 99.29% случаев изменения х приводят к изменению y. Другими словами - точность подбора уравнения регрессии - высокая.

Остальные 0.71% изменения Y объясняются факторами, не учтенными в модели (а также ошибками спецификации)

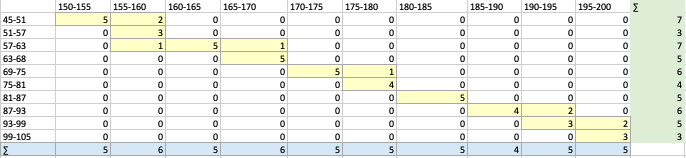


= Yi – b \*

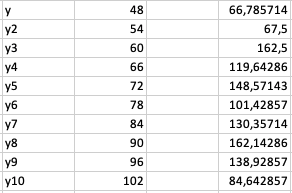
0,017597

1. переход к коррел. таблице



8)





1. Выводы

Изучена зависимость Y от X. На этапе спецификации была выбрана парная линейная регрессия.

Оценены её параметры методом наименьших квадратов. Статистическая значимость уравнения проверена с помощью коэффициента детерминации и критерия Фишера.

Установлено, что в исследуемой ситуации 99.86% общей вариабельности Y объясняется изменением X.

Установлено также, что параметры модели статистически значимы.

Возможна экономическая интерпретация параметров модели - увеличение X на 1 ед.изм. приводит к увеличению Y в среднем на 1.172 ед.изм.

Полученные оценки уравнения регрессии позволяют использовать его для прогноза.

При x=0.05, Y будет находиться в пределах от -133.51 до -128.34 ед.изм. и с вероятностью 95% не выйдет за эти пределы

Пример прогноза:  
выберем пару x , y , подставим в ур-е регрессии :

Y = 204 , X = 105 => 204 = 1,172t – 130,98 , t = -154,8