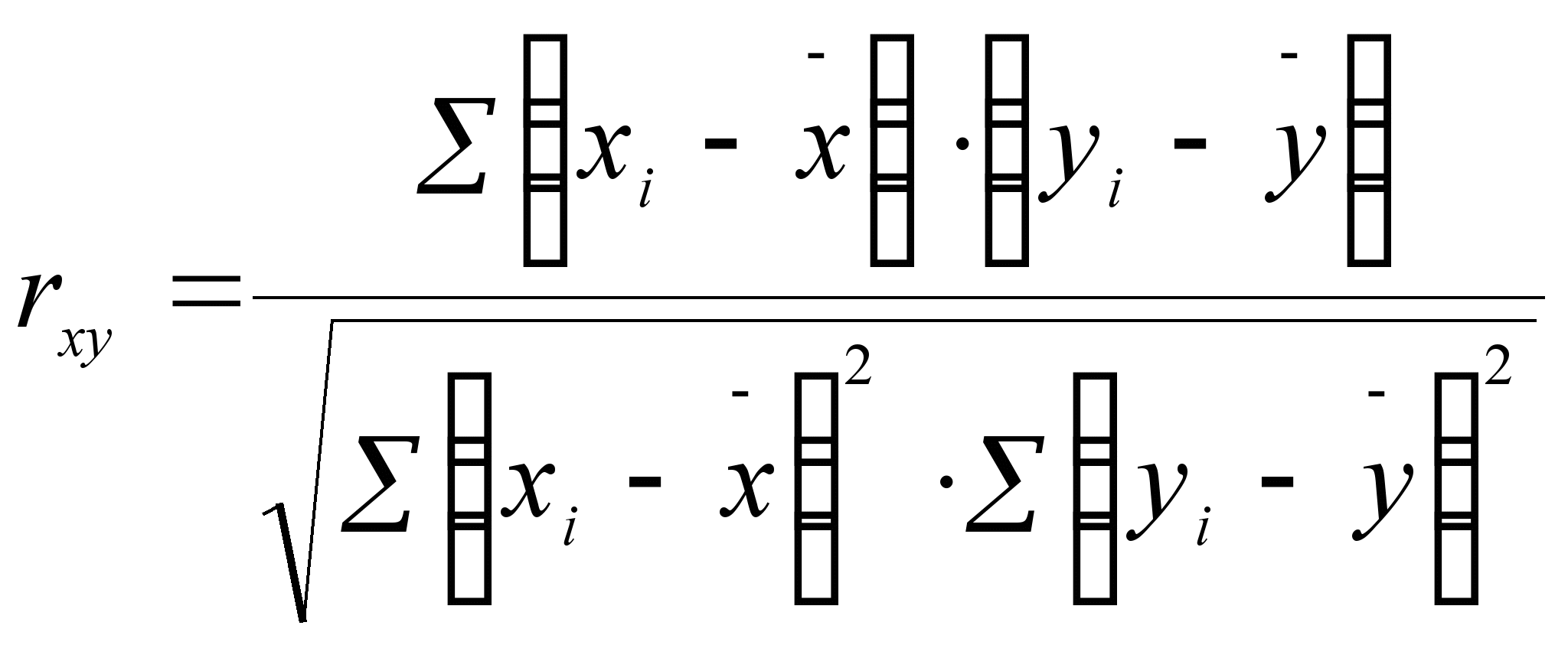
Минский филиал МЭСИ Профессионально-ориентированные экономические информационные системы

КОРРЕЛЯЦИОННЫЙ И РЕГРЕССИОННЫЙ АНАЛИЗ В MS EXCEL И В СТАТИСТИЧЕСКОМ КОМПЛЕКСЕ SPPS.  
  
Оглавление.

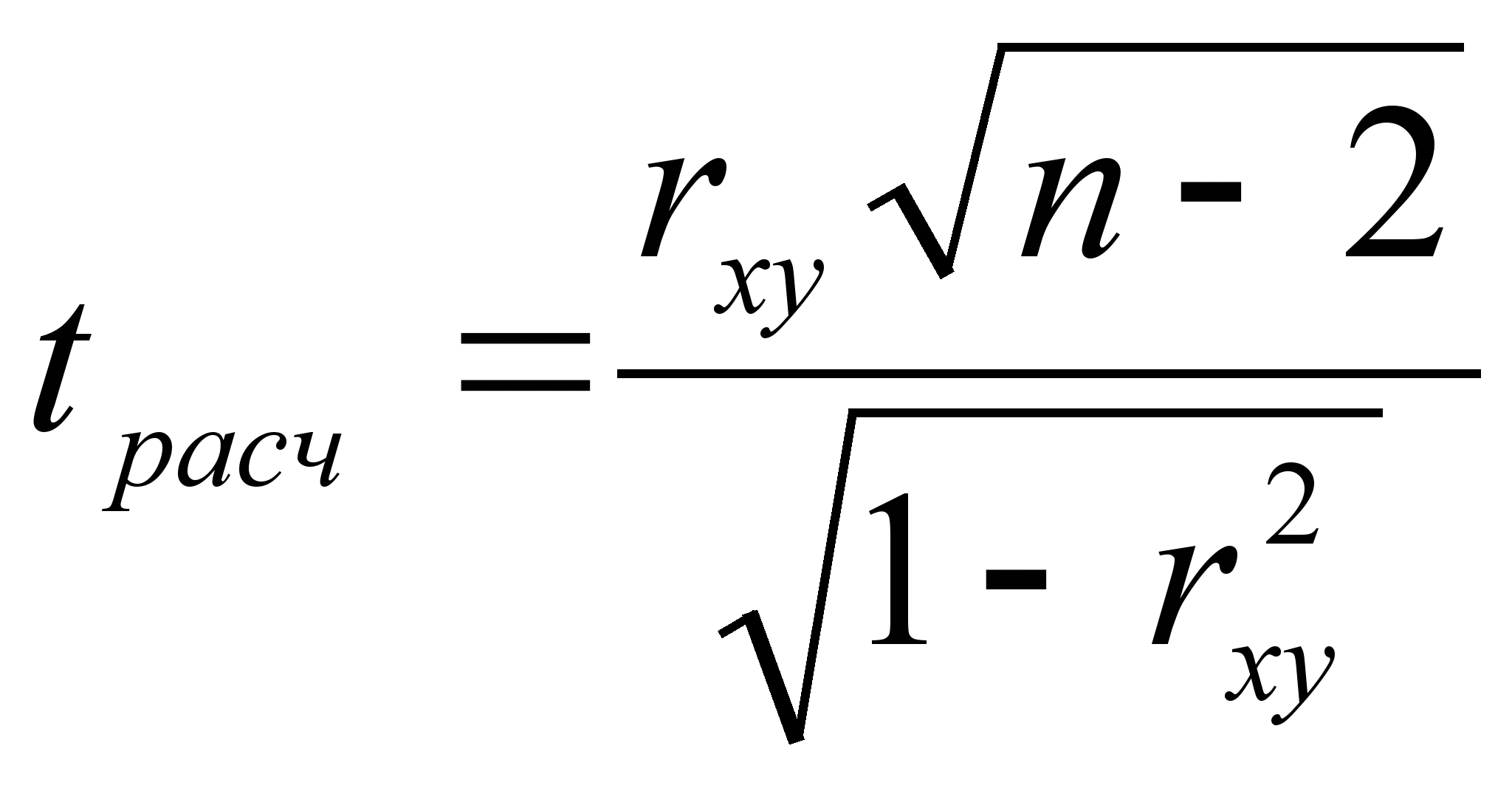
1. Параметрический корреляционный анализ. 3  
  
Задание 1. Проведение параметрического корреляционного анализа. 5  
  
1.1 Расчет коэффициента корреляции Пирсона с использованием формулы (1). 5  
  
1.2 Расчет коэффициента корреляции с использованием функции КОРРЕЛ. 6  
  
1.3 Расчет коэффициента корреляции с использованием Пакета анализа MS Excel. 6  
  
1.4 Проведение параметрического корреляционного анализа в ППП SPSS. 8  
  
2. Непараметрические методы анализа. 12  
  
Задание 2. Использование непараметрических методов анализа. 14  
  
2.1 Расчет коэффициента корреляции Спирмена с использованием формулы (3). 14  
  
2.2 Расчет коэффициента корреляции Кенделла с использованием формулы (4). 15  
  
2.3 Расчет коэффициента корреляции Спирмена и Кенделла в ППП SPSS. 16  
  
Задание для самостоятельной работы. 17  
  
3. Регрессионный анализ. 18  
  
3.1 Однопараметрические модели. 18  
  
Задание 3. Построение модели линейной регрессии. 19  
  
3.1 Расчет параметров уравнения линейной регрессии с использованием функции ЛИНЕЙН. 19  
  
3.2 Нахождение уравнения линейной регрессии графическим методом. 21  
  
3.3 Построение модели линейной регрессии с помощью инструмента «Регрессия». 21  
  
3.4 Построение модели линейной регрессии с помощью ППП SPSS. 24  
  
3.5. Построение линии регрессии. 26

^

**1. Параметрический корреляционный анализ.**

Одна из наиболее распространенных задач статистического исследования состоит в изучении **связи между выборками**. Обычно связь между выборками носит не функциональный, а вероятностный (или стохастический) характер. В этом случае нет строгой, однозначной зависимости между величинами. При изучении стохастических зависимостей разли­чают **корреляцию и регрессию**.  
  
**^ Корреляционный анализ** состоит в определении степени связи между двумя случайными величинами *X*и *Y.*В качестве меры такой связи используется **коэффи­циент корреляции.**Коэффициент корреляции оценивается по выборке объема *п*связанных пар наблюдений *(xi, yi*) из совместной генеральной совокупности *X*и *Y.*Существует несколько типов коэффициентов корреляции, применение которых зависит от измерения (способа шкалирования) величин *X*и *Y.*  
  
Для оценки степени взаимосвязи величин *X*и *Y,*измеренных в количественных шкалах, используется коэффициент линейной корреляции (**коэффициент Пирсона**), предполагающий, что выборки *X*и *Y* распределены **по нормальному закону.**При таком распределении большая часть значений группируется около некоторого среднего значения, по обе стороны от которого частота наблюдений равномерно снижается.  
  
**^ 1. Линейный коэффициент корреляции***—*параметр, который характеризует степень линейной взаимосвязи между двумя выборками, рассчитывается по формуле (1):  
  
 (1)  
  
где *хi* — значения, принимаемые в выборке X,   
  
*yi* — значения, принимаемые в выборке Y;   
  
http://www.studmed.ru/docs/static/9/f/2/e/f/9f2ef2d0303.png — средняя по X, http://www.studmed.ru/docs/static/2/1/7/9/6/217964d8ab5.png — средняя по Y.   
  
Коэффициент корреляции изменяется от -1 до 1. Когда при расчете получается величина большая 1 или меньшая -1 — следовательно, произошла ошибка в вычислениях. При значении 0 линейной зависимости между двумя выборками нет.   
  
Знак коэффициента корреляции очень важен для интерпре­тации полученной связи. Если знак коэффициента линейной корреляции — **плюс**, то связь между коррелирующими признаками такова, что большей величине одного признака (переменной) соответствует большая величина другого признака (другой переменной). Иными словами, если один показатель (переменная) увеличивается, то соответственно увеличивается и другой показатель (переменная). Такая зависимость носит название **прямо пропорциональной зависимости.**  
  
Если же получен знак **минус**, то большей величине одного признака соответствует меньшая величина другого. Иначе говоря, при наличии знака минус, увеличению одной переменной (признака, значения) соответствует уменьшение другой переменной. Такая зависимость носит название **обратно пропорциональной зависимости.**  
  
**Теснота связи и величина коэффициента корреляции.**

|  |  |
| --- | --- |
| **^ Коэффициент корреляции rxy** | **Теснота связи** |
| 0,91-1,0 | Очень сильная |
| 0,81-0,9 | Весьма сильная |
| 0,65-0,8 | Сильная |
| 0,45-0,64 | Умеренная |
| 0,25-0,44 | Слабая |
| До 0,25 | Очень слабая |
| **« » - прямая зависимость**  **«-» - обратная зависимость** |  |

Для того чтобы оценить наличие связи между двумя переменными, также можно использовать **t-статистику Стьюдента**, которая оценивает отношение величины линейного коэффициента корреляции к среднему квадратическому отклонению и рассчитывается по формуле (2)  
  
 (2)  
  
Полученную величину tрасчсравнивают с табличным значением t-критерия Стьюдента с n-2 степенями свободы. Если tрасч> tтабл,то практически невероятно, что найденное значение обусловлено только случайными совпадениями величин X и Y d в выборке из генеральной совокупности, т.е. существует зависимость между X и Y. И наоборот, если tрасч< tтабл, то величины X и Y независимы.  
  
^

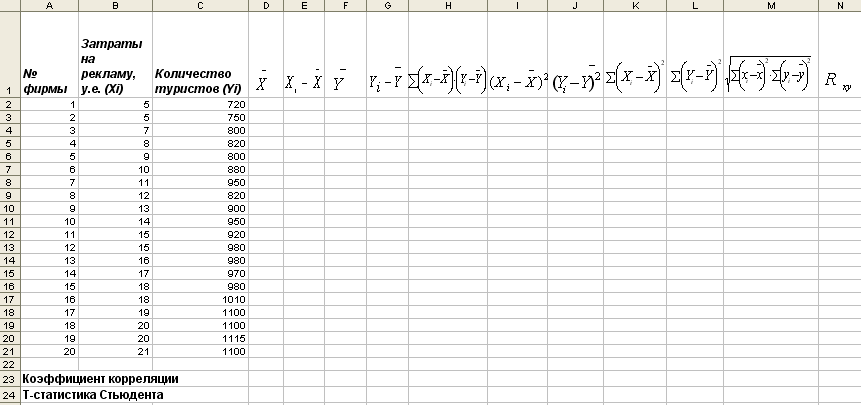
**Задание 1. Проведение параметрического корреляционного анализа.**

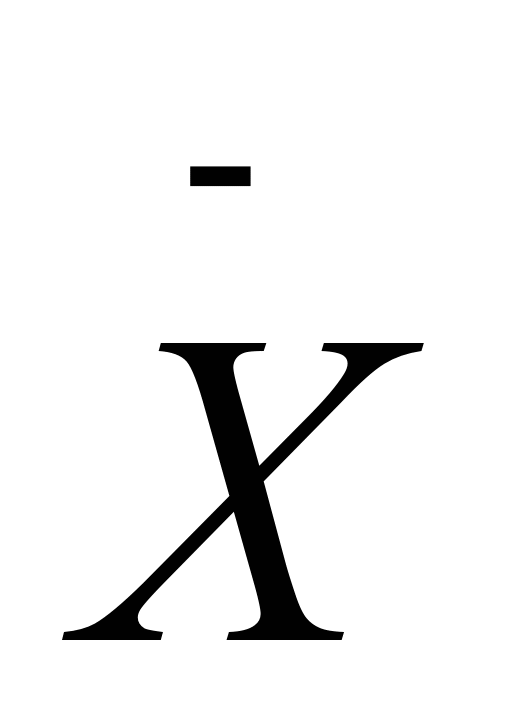
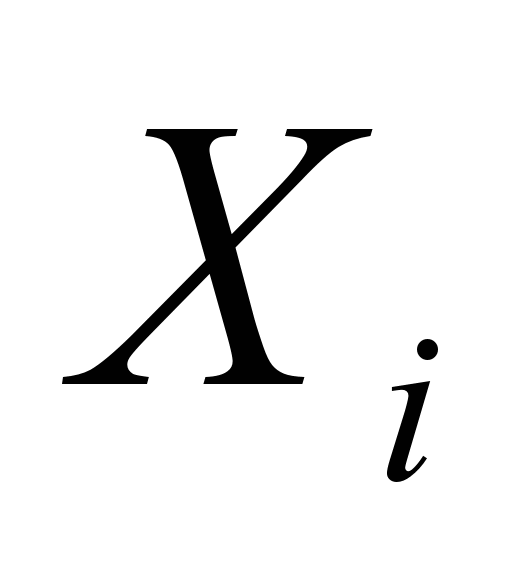
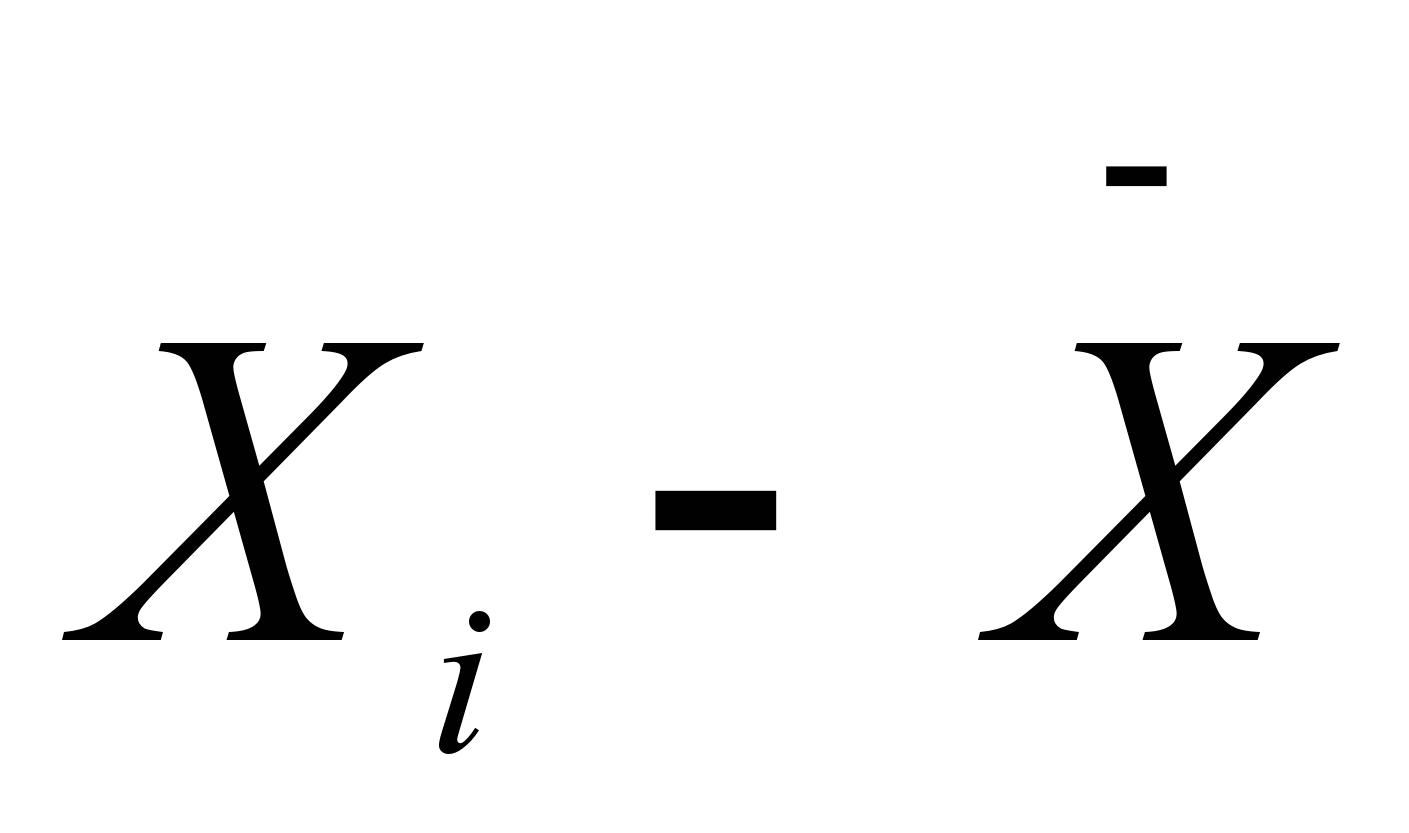
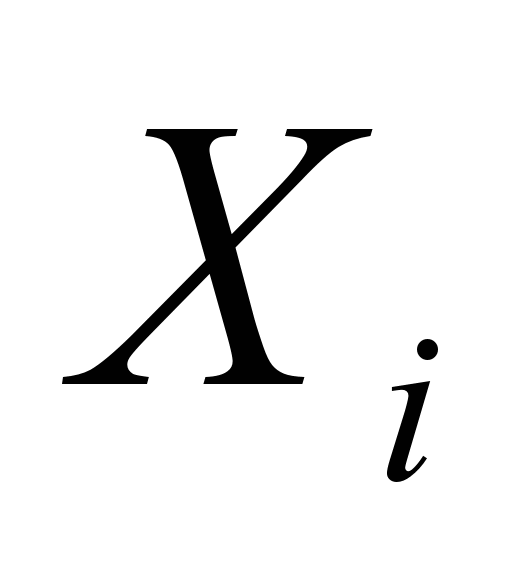
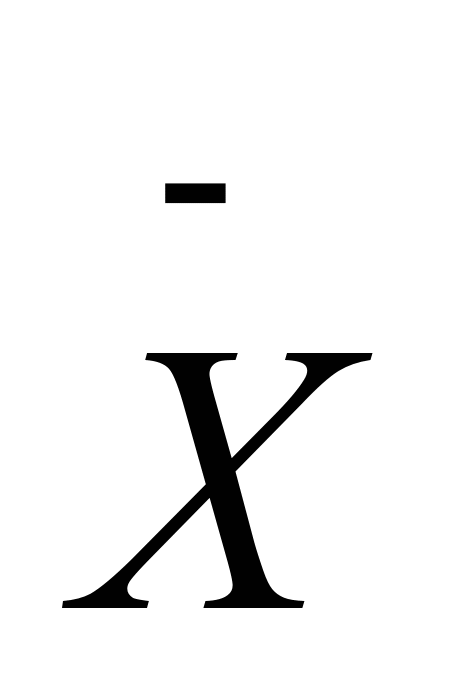
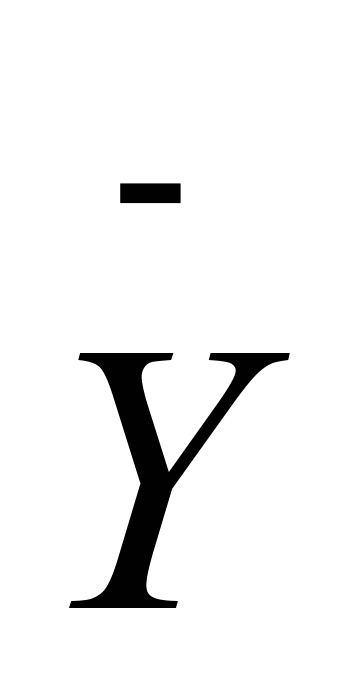
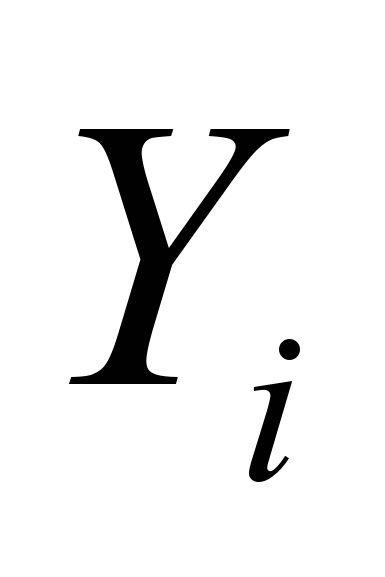
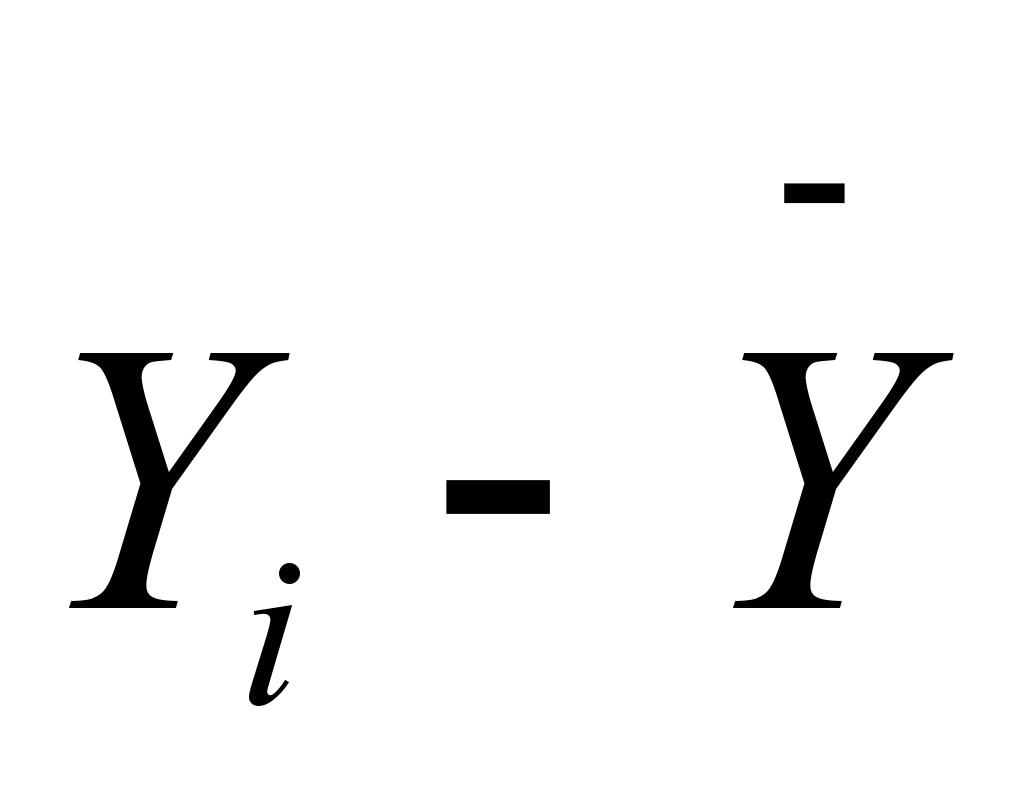
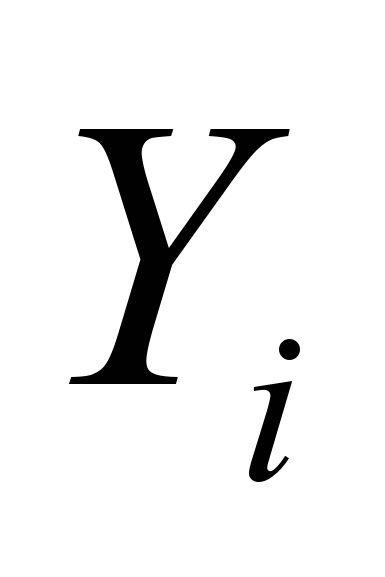
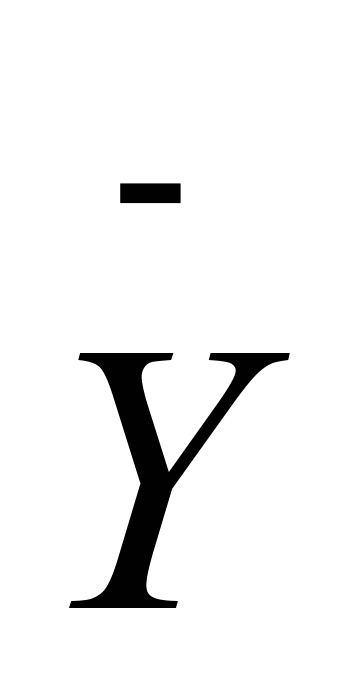
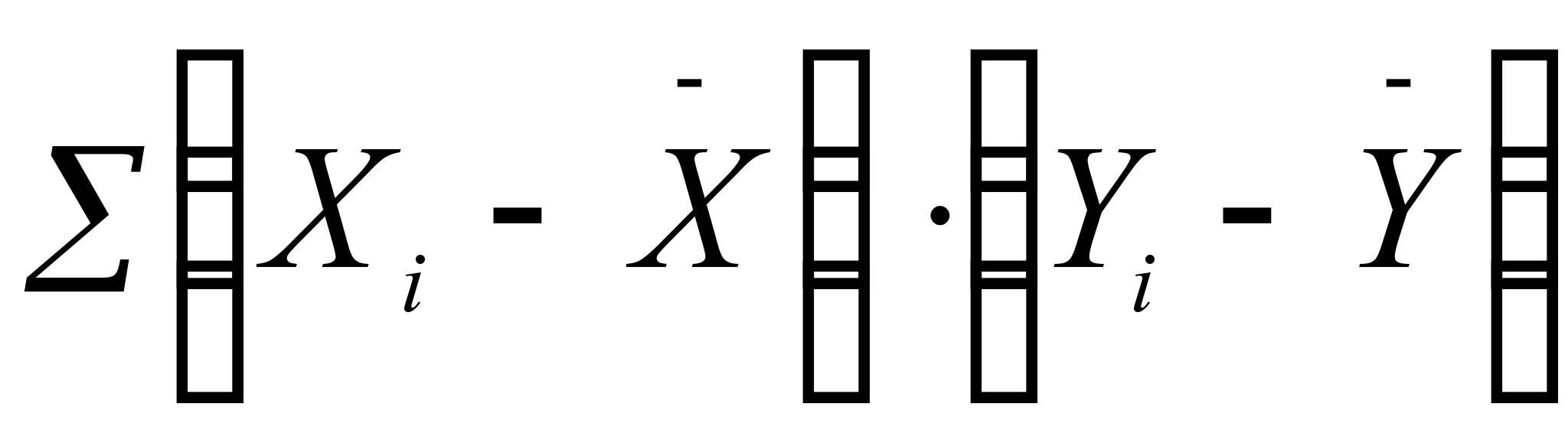
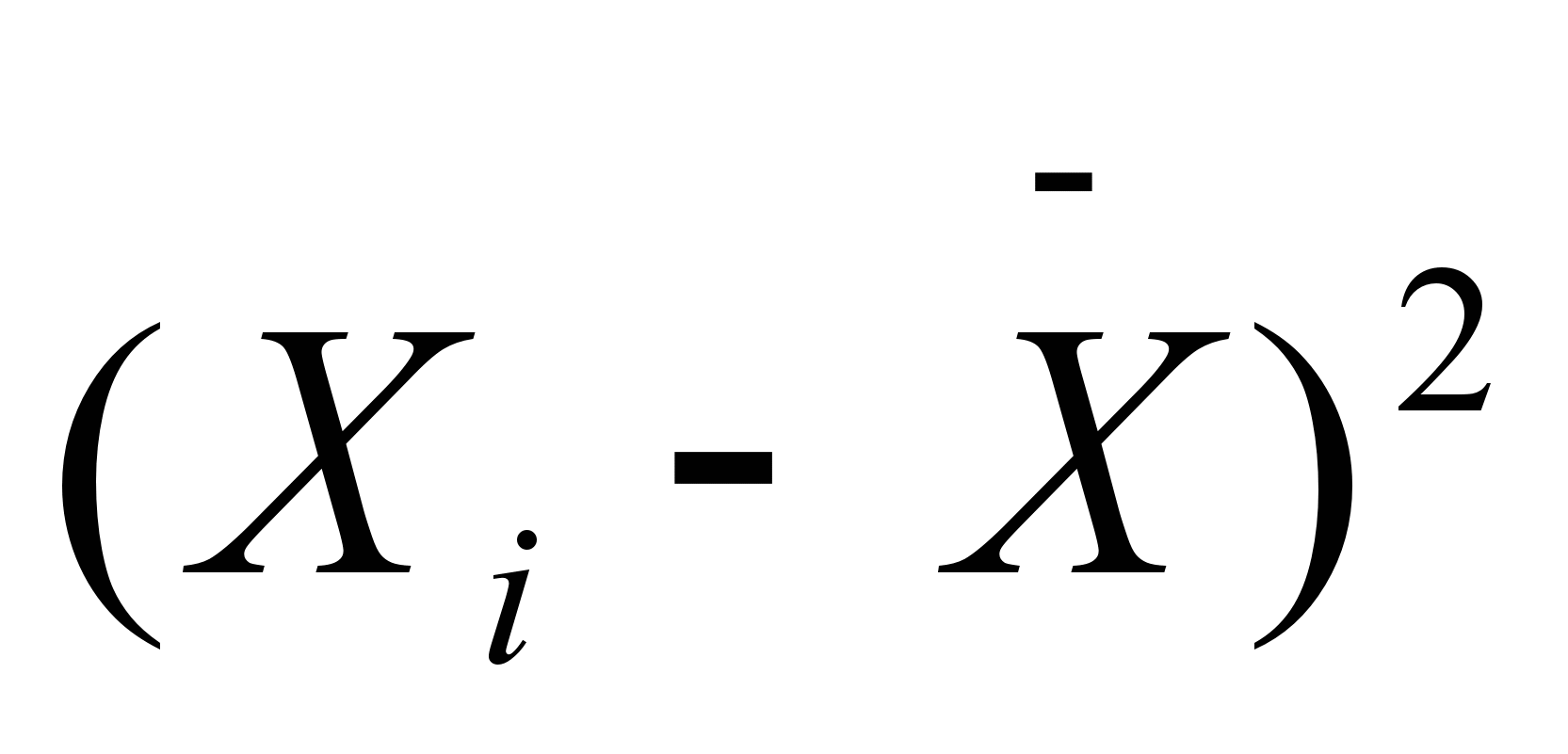
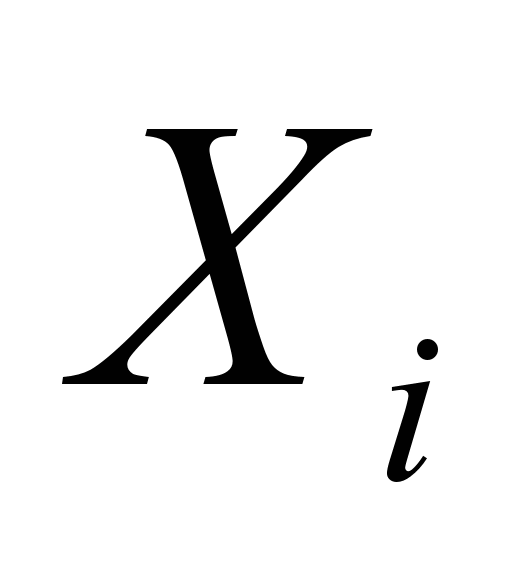
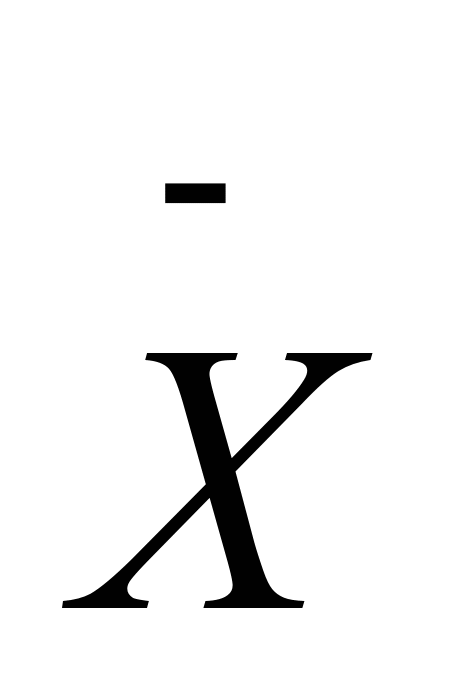
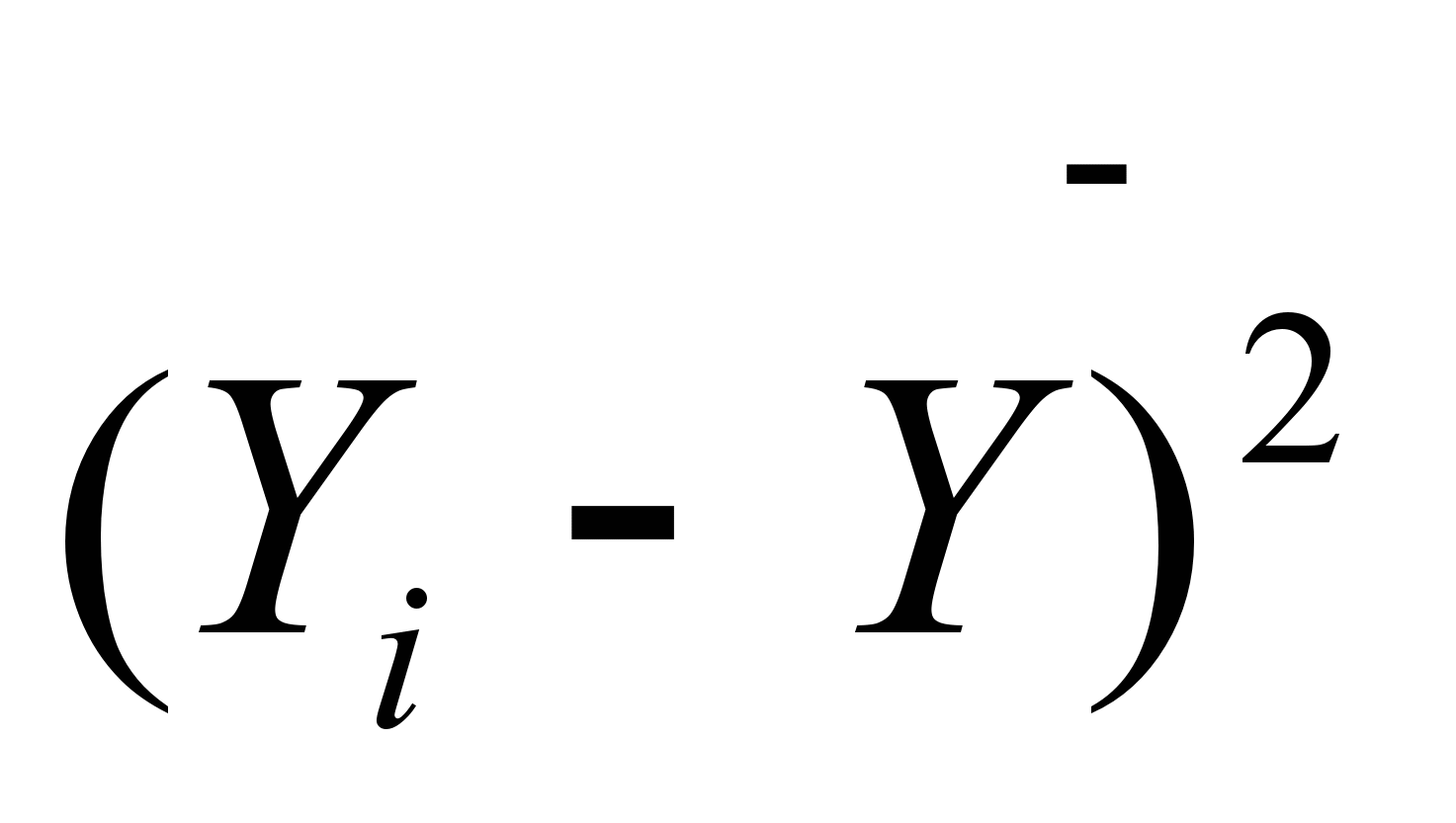
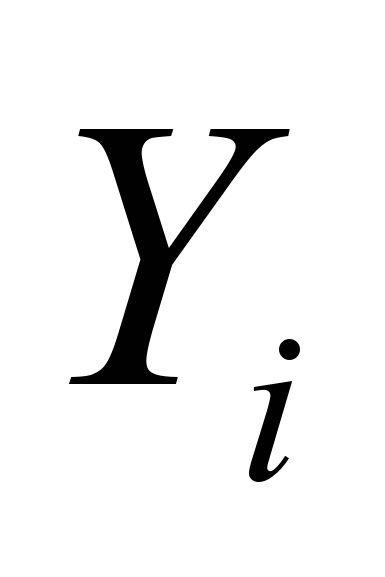
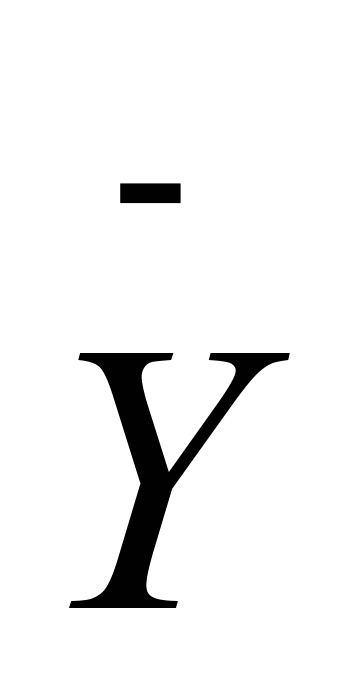
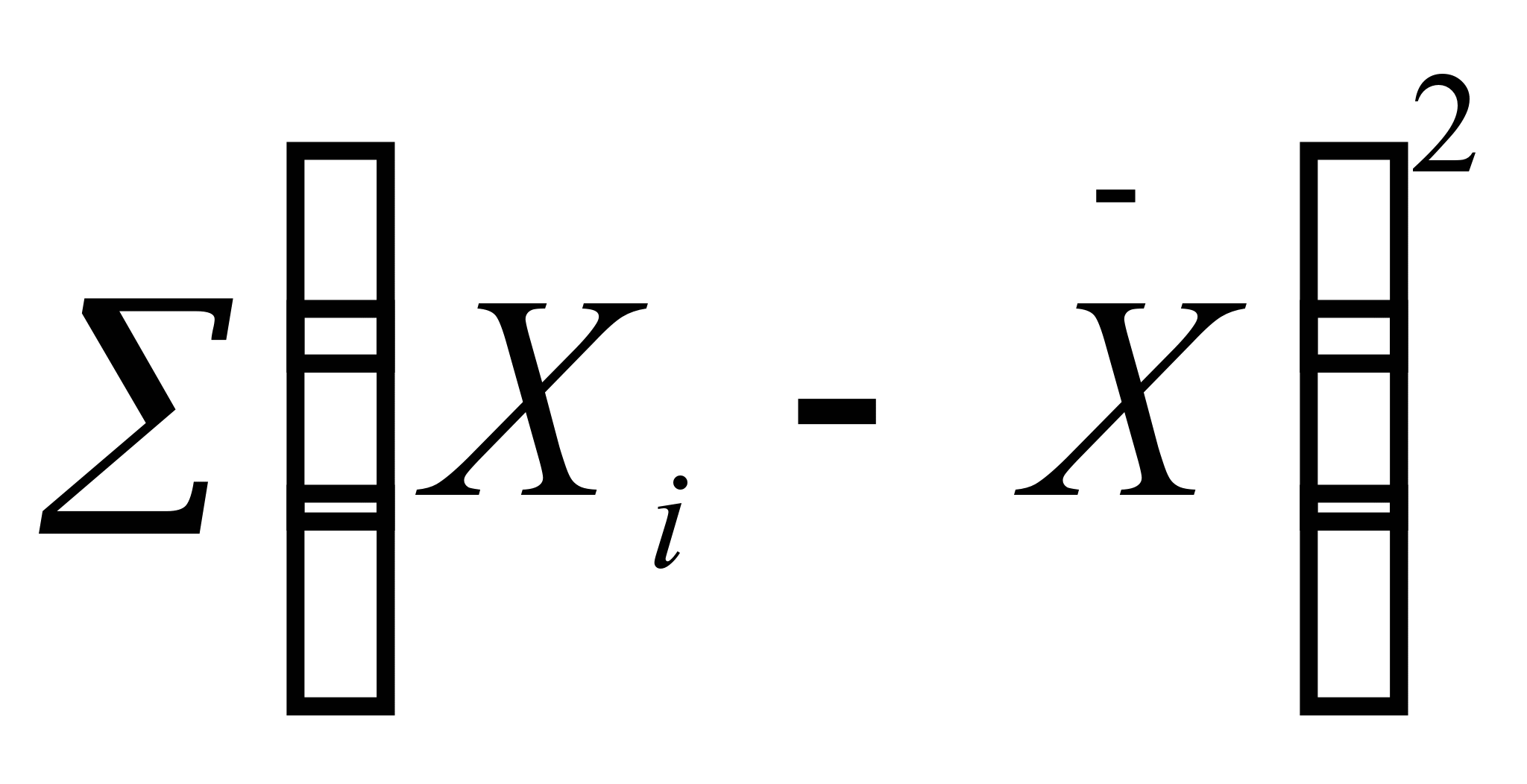
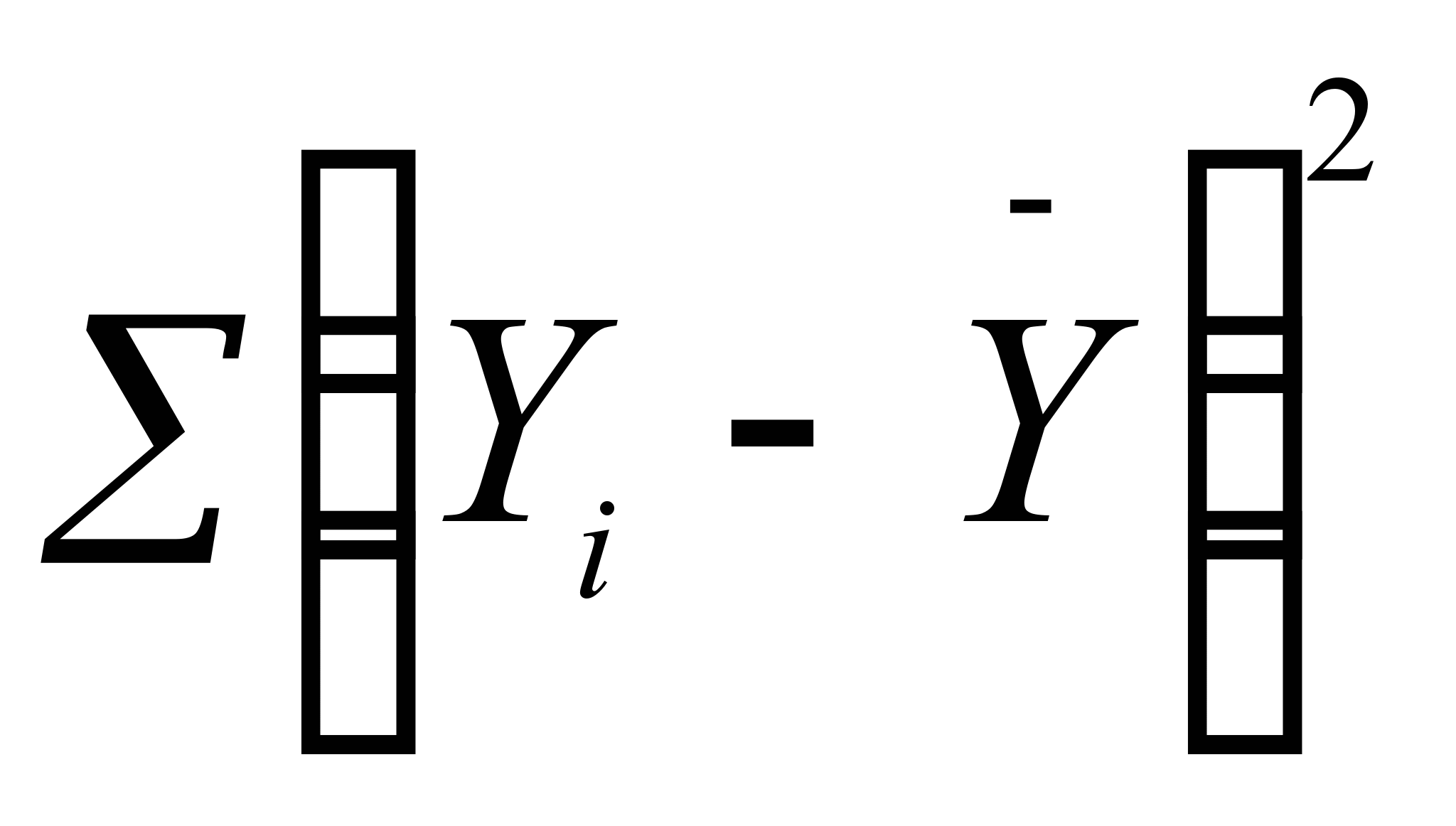
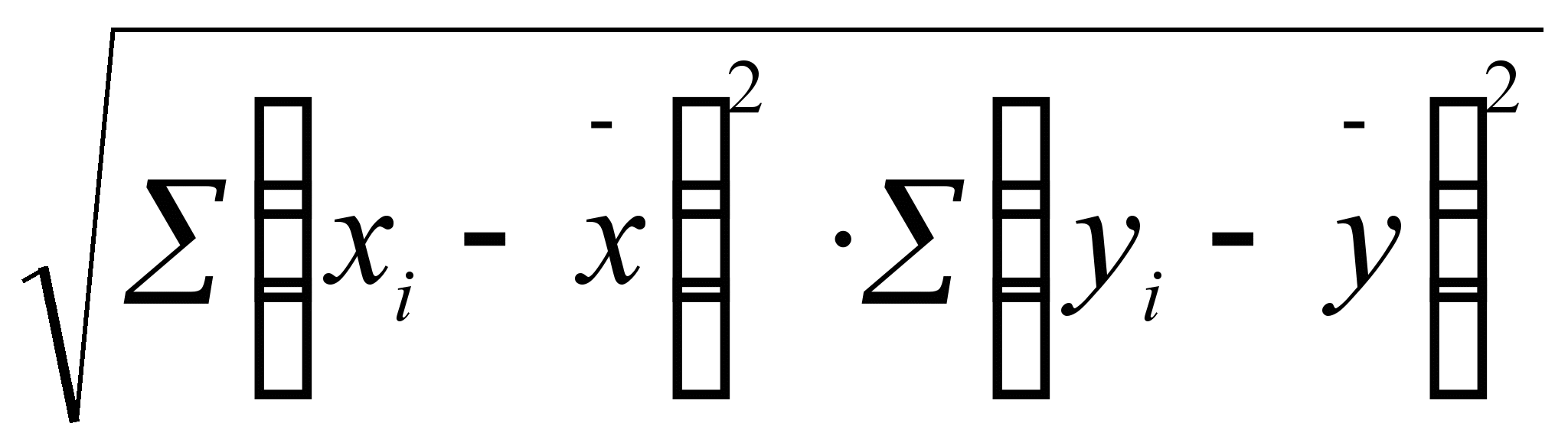
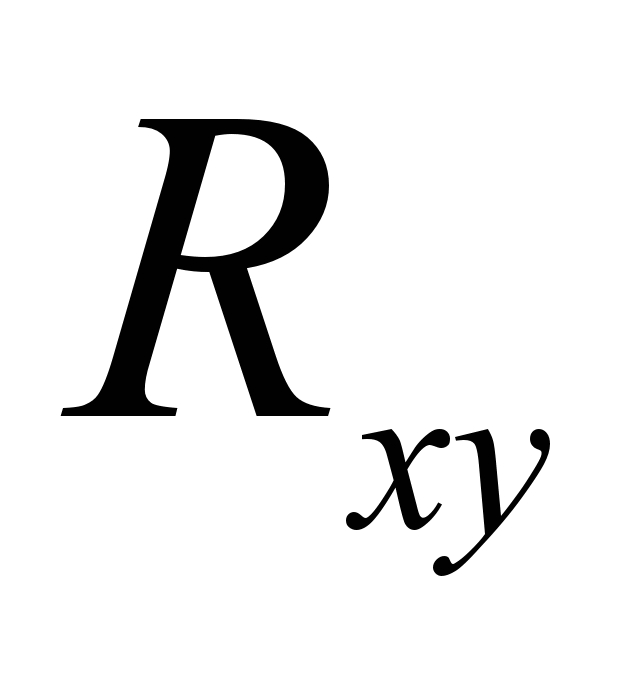
**Условие задачи:***По 20 туристическим фирмам были установлены затраты на рекламную кампанию и количество туристов, воспользовавшихся после ее проведения услугами каждой фирмы. Необходимо определить коэффициент корреляции между исследуемыми признаками, используя:*

1. *Формулу линейного коэффициента корреляции.*
2. *Статистическую функцию КОРРЕЛ.*
3. *Пакет анализа MS Excel.*
4. *Методы корреляционного анализа ППП SPSS.*

^

**1.1 Расчет коэффициента корреляции Пирсона с использованием формулы (1).**

Откройте новую книгу MS Excel и создайте таблицу согласно рис. 1, сохраните в своей папке под именем *Параметрический\_Анализ.xls*.  
  
  
  
**Рис.1. Данные для задачи №1.**

1. В ячейке **D2** рассчитайте  (среднее значение переменной ).
2. В диапазоне ячеек **E2:E21** рассчитайте (отклонение значений переменной от среднего значения ).
3. В ячейке **F2** рассчитайте  (среднее значение переменной ).
4. В диапазоне ячеек **G2:G21** рассчитайте (отклонение значений переменной от среднего значения).
5. В ячейке **H2** рассчитайте сумму произведений , воспользовавшись математической функцией СУММПРОИЗВ.
6. В диапазоне ячеек **I2:I21** рассчитайте (квадраты отклонений значений переменной от среднего значения).
7. В диапазоне ячеек **J2:J21** рассчитайте (квадраты отклонений значений переменной от среднего значения).
8. В ячейке **K2** рассчитайте .
9. В ячейке **L2** рассчитайте .
10. В ячейке **M2** рассчитайте .
11. В ячейке **N2** рассчитайте коэффициент линейной корреляции .
12. Сделайте вывод о тесноте связи между затратами на рекламу и количеством привлеченных туристов и направлении этой связи.

^

**1.2 Расчет коэффициента корреляции с использованием функции КОРРЕЛ.**

1. В ячейке **D23** рассчитайте коэффициент корреляции, используя функцию **КОРРЕЛ**из категории **Статистические.**

|  |
| --- |
| Синтаксис функции **^ КОРРЕЛ:**  **КОРРЕЛ (массив 1; массив 2):**  где *массив 1* – ссылка на диапазон ячеек первой выборки (X);  *массив 2* – ссылка на диапазон ячеек второй выборки (Y). |

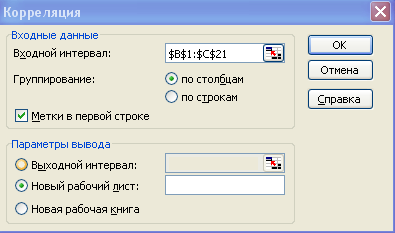
http://www.studmed.ru/docs/static/0/6/8/2/d/0682d9dd855.png

1. Оцените значимость коэффициента корреляции. С этой целью рассматриваются две гипотезы. **Основная Н0**ρ: xy=0 и **альтернативная Н1**ρ: xy≠0. Для проверки гипотезы Н0 рассчитайте **t-статистику Стьюдента**поформуле (2) в ячейке **D24.** В нашем случае число степеней свободы ν = n-2=20-2 = 18 (количество наблюдений минус два).
2. Сравните полученное значение с критическим значением tν,α распределения Стьюдента. (При ν =18 и вероятности α = 0,05, tν,α,табл= 1,734). Сделайте вывод о значимости коэффициента корреляции.

http://www.studmed.ru/docs/static/c/6/0/1/e/c601e27f3b4.png  
^

**1.3 Расчет коэффициента корреляции с использованием Пакета анализа MS Excel.**

1. Сначала убедитесь, что был активизирован **Пакет анализа,**т.е. в меню **Сервис** есть команда **Анализ данных**. Если нет, то выполните команду **Сервис/Надстройки.**В диалоговом окне **Надстройки**установите флажок **Пакет анализа** и щелкните по кнопке **ОК.**
2. Далее выполните команду **Сервис/Анализ данных.** Выберите инструмент анализа **Корреляции**.
3. В диалоговом окне **Корреляция**установите параметры согласно рис.2.

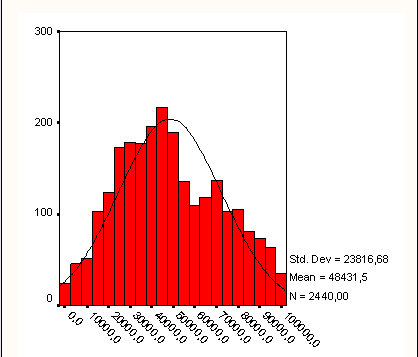
  
  
**Рис.2. Окно Корреляция.**

1. На новом рабочем листе получите значение коэффициента корреляции. Сравните с ранее полученными результатами.
2. Сохраните изменения в книге.

^

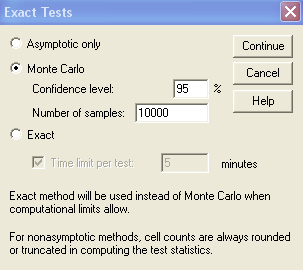
**1.4 Проведение параметрического корреляционного анализа в ППП SPSS.**

1. Загрузите SPSS и введите данные в режиме **Data View (Просмотр данных)** согласно рис. 3. Сохраните в своей папке под именем *Параметрический\_Анализ. Sav (*данные можно также скопировать из файла MS Excel).

  
  
**Рис.3. Данные в режиме** *Data View.*  
Согласно статистической теории, чтобы сделать возможным применение большинства статистических процедур, данные должны подчиняться **закону нормального распределения,**при котором большая часть значений группируется около некоторого среднего значения, по обе стороны от которого частота наблюдений равномерно снижается. На диаграмме (рис. 4) показана кривая нормального распределения (Колокол Гаусса).   
  
  
  
**^ Рис.4. Кривая нормального распределения.**  
  
Для того, чтобы определить подчиняются ли два эмпирических распределения одному закону, либо определить, подчиняется ли полученное распределение предполагаемой модели используется [**критерий согласия**](http://ru.wikipedia.org/wiki/%D1%81%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B8%D0%B9_%D0%BA%D1%80%D0%B8%D1%82%D0%B5%D1%80%D0%B8%D0%B9)**Колмогорова**(также известный, как критерий согласия Колмогорова-Смирнова). 

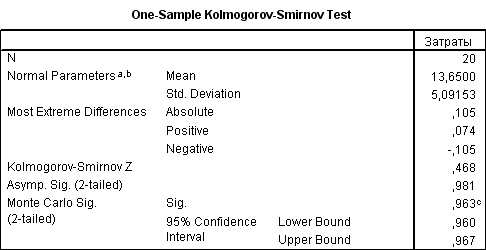
1. Прежде чем исследовать наличие корреляции между исследуемыми признаками протестируем обе переменные **^ Затраты** и **Количество туристов** на закон нормального распределения при помощи критерия согласия Колмогорова-Смирнова. Для этого:

* Выполните команду **Analyze/Nonparametric Tests/1-Sample K-S***(Анализ/Непараметрические тесты/Тест Колмогорова-Смирнова).*
* В диалоговом окне перетащите переменную **Затраты** в область **Test Variable List***(Список тестируемых переменных).*
* Выберите тип распределения **Normal***(Нормальное распределение).*
* При малых размерах выборки применяются точные методы. Выберите использование точного метода при помощи кнопки **[Exact]***(Точный).* Для практических целей следует применять метод **Monte-Carlo** с установленным по умолчанию количеством выбо­рок (10 000). Доверительный уровень 99 % практически всегда является слишком высоким, поэтому измените его на 95 %, что соответствует доверительному уров­ню при расчете статистической ошибки выборки для маркетинговых исследова­ний (рис.5).

  
  
**Рис.5. Настройка параметров точного метода.**

* *Нажмите кнопку [Continue] (Продолжить) и ОК.*

1. Проанализируем полученный результат (см. рис.6).

  
  
**Рис.6. Результат тестирования на закон нормального распределения.**

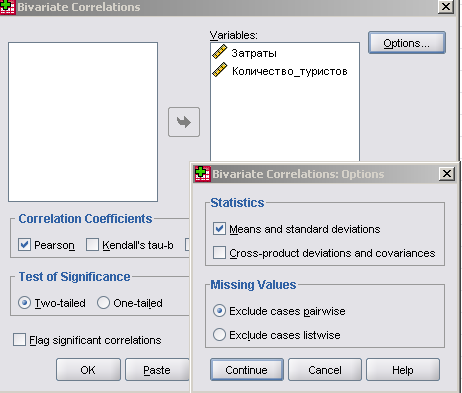
* **Mean***(среднее значение)*=13,65 у.е;
* **Std. deviation** *(Стандартное отклонение)* = 5,091153 у.е, это означает, что в интервале шириной, равной удвоенному стандартному отклонению, который отложен по обе стороны от среднего значения, располагается примерно 67% всех значений выборки, подчиняющейся нормальному распределению).
* **Значимость**тестовых характеристик:**Asymp. Sig. (2-tai-led)** (*асимптотическая значимость двухсторонняя*) = 0,981 и **Monte Carlo Sig***(точная значимость Монте-Карло)* = 0,963. Поскольку исходная гипотеза состояла в наличии нормального распределения, то статистическая значимость менее 0,05 означает, что исследуемая переменная не подчиняется закону нормального распределения. В нашем случае статистически значимого отличия от нормального распределения не обнаружено **p=0,963≥0,05**иисходная гипотеза подтверждается.
* Для визуального просмотра функции нормального распределения выполните команду **Graphs/Histogram***(Графики/Гистограмма).*В диалоговом окне в область**Variable***(Переменная)*перетащите переменную **Затраты**и установите флажок**Display normal curve***(Отображать нормальную кривую).*

1. Аналогичным образом протестируйте другую переменную **^ Количество туристов** на закон нормального распределения.

Следует отметить, что в маркетинговых исследованиях данные редко оказываются подчиненными закону нормального распределения. Многие аналитики предполагают, что данные, не подчиняющиеся нормальному распределению, являются выбросами (случайными значениями). Данная техника оправдывает себя в тех случаях, когда от абсолютной точности построенных статистических моделей ровным счетом ничего не зависит. Исследователей в большинстве случаев интересует лишь общее направление различий, связей и т. п. В этом и заключается специфика маркетинговых исследований: аналитика не интересует, как ведет себя каждый респондент в выборке, — ему интересно знать, как ведут себя целевые группы.

1. После того как переменные протестированы на закон нормального распределения, можно приступить к определению линейного коэффициента корреляции Пирсона. Для этого:

* Выполните команду **Analyze/Correlate/Bivariate** *(Анализ/Корреляция/Парные).*
* В диалоговом окне в область переменных перетащите обе переменные, укажите коэффициент корреляции Пирсона, уровень значимости - двусторонний, нажмите кнопку **[Options]** и установите параметры согласно рис. 7.
* Нажмите кнопку **[Continue]***(Продолжить)* и **ОК**.

  
  
**Рис.7. Диалоговое окно «Корреляции».**

1. Проанализируем полученный результат (см. рис.8).

* + В верхней области отчета SPSS вывел таблицу **Descriptive Statistics** (*Описательная статистика*), поскольку был установлен флажок **Means and standard deviations** *(Среднее значение и стандартное отклонение).*
  + Из таблицы **Correlations***(Корреляции)* видно, что коэффициент корреляции Пирсона равен0,952**,** это указывает на то, что зависимость между затратами на рекламу и количеством привлеченных туристов прямо пропорциональная, а **теснота связи очень сильная.**
  + Статистическая значимость (**Sig. (2-tailed)**) в нашем случае меньше 0,001, показывает вероятность того, что корреляции не является случайной, т.е. с вероятностью 95% можно утверждать, что коэффициент корреляции является значимым.

  
  
**Рис.8. Отчет по корреляции Пирсона.**

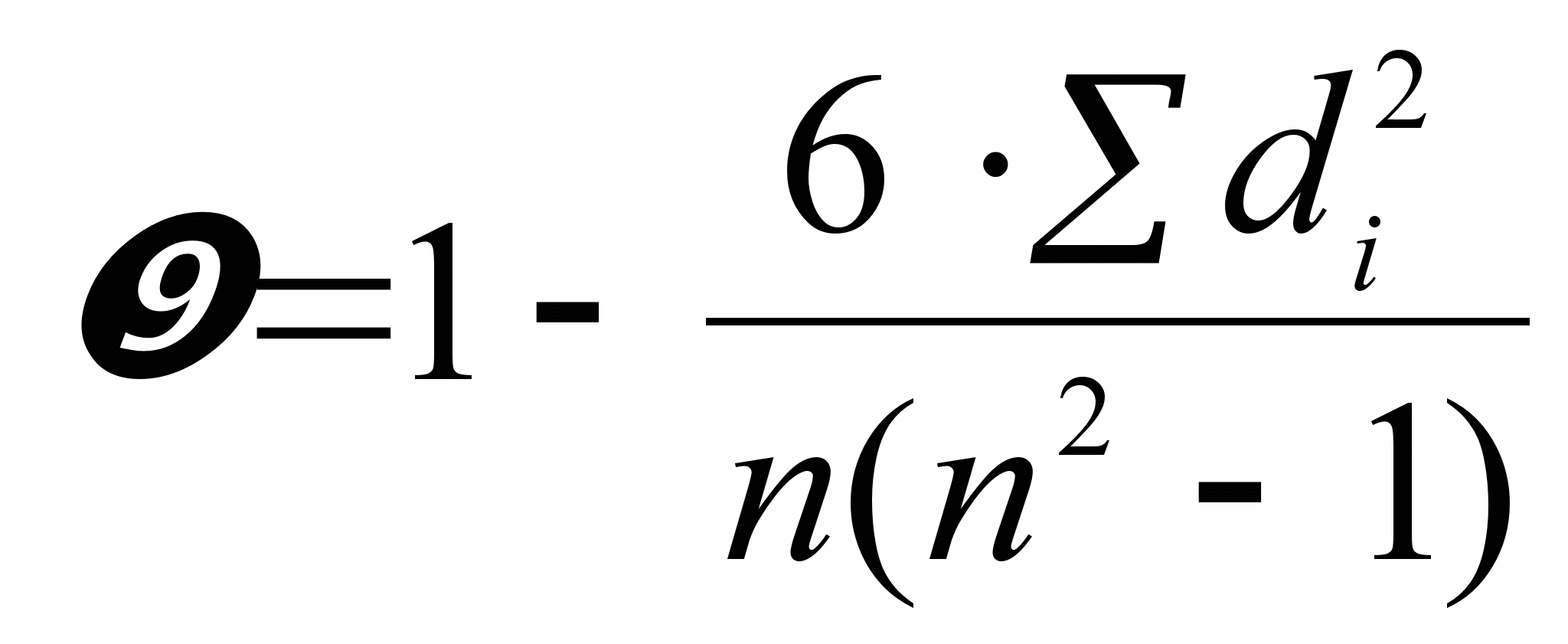
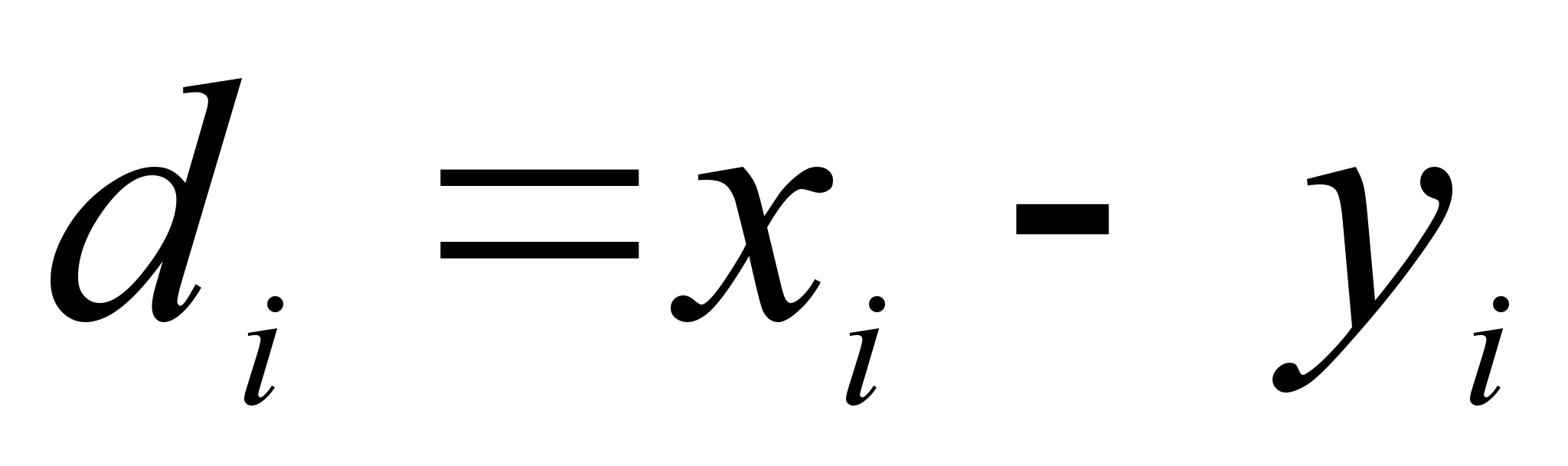
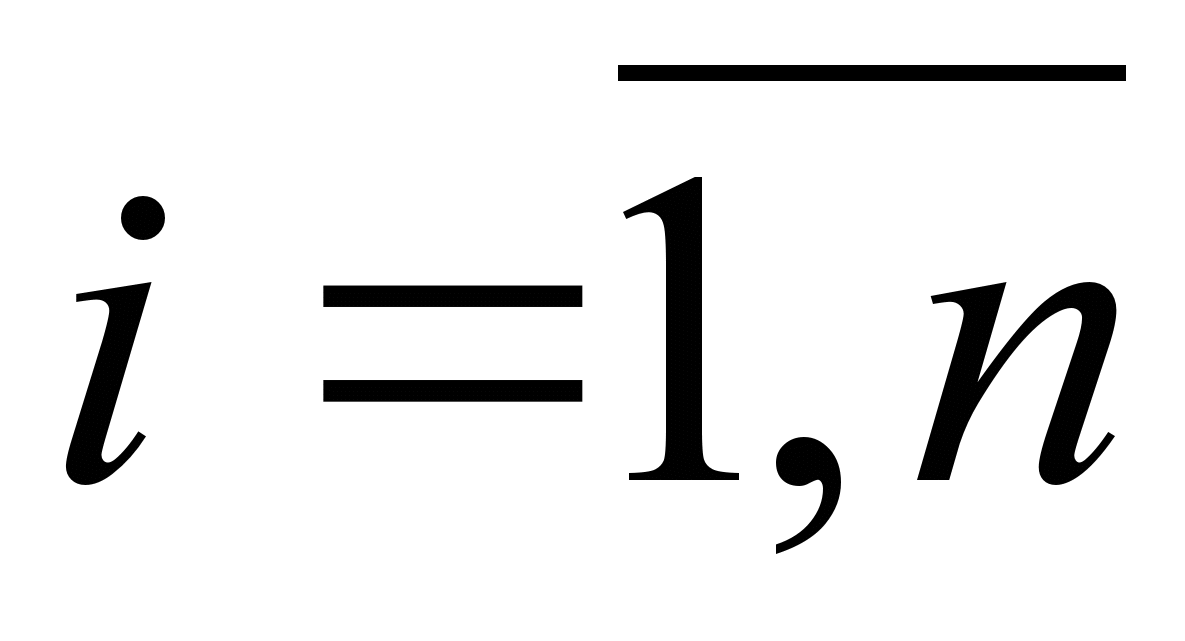
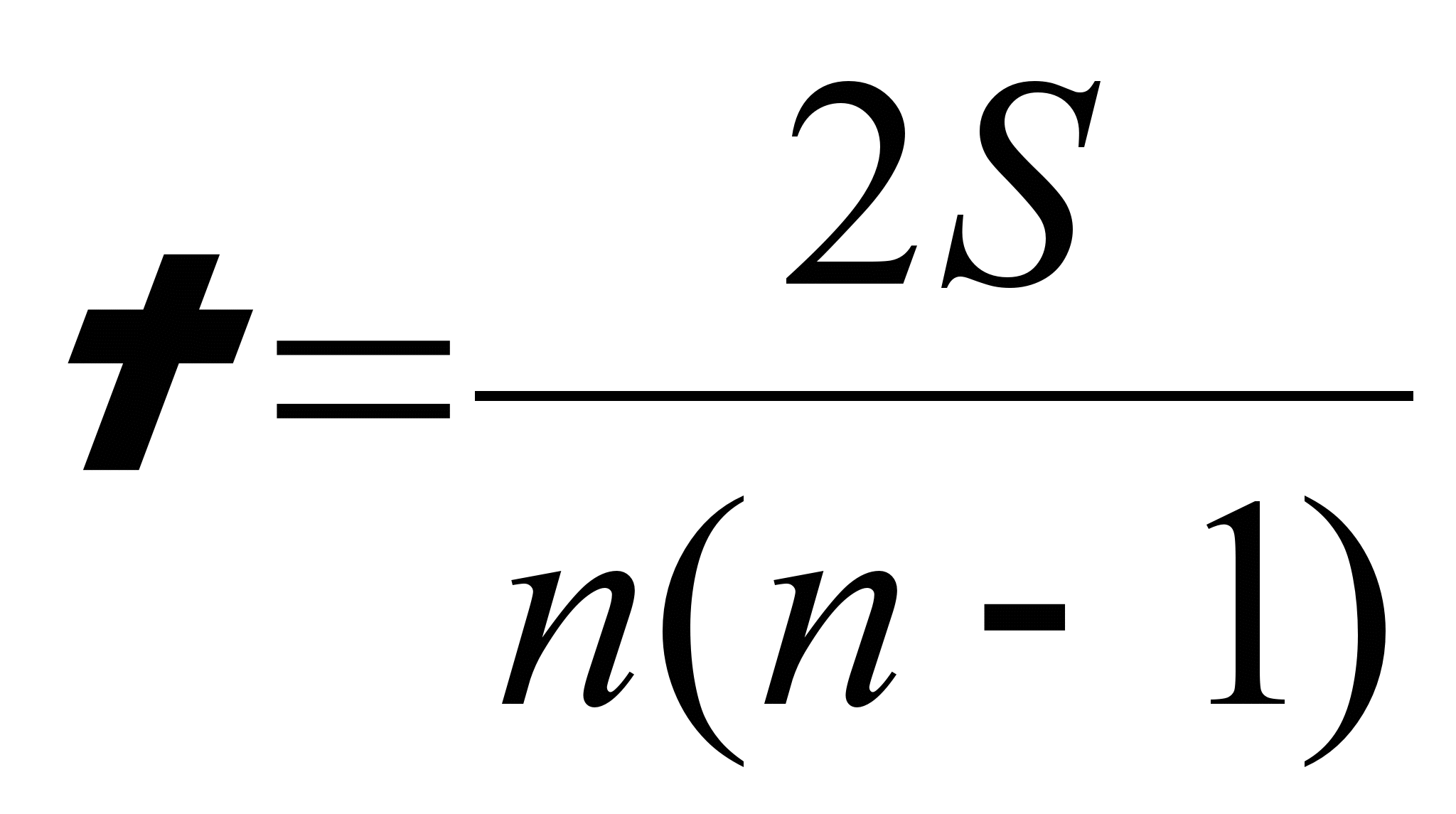
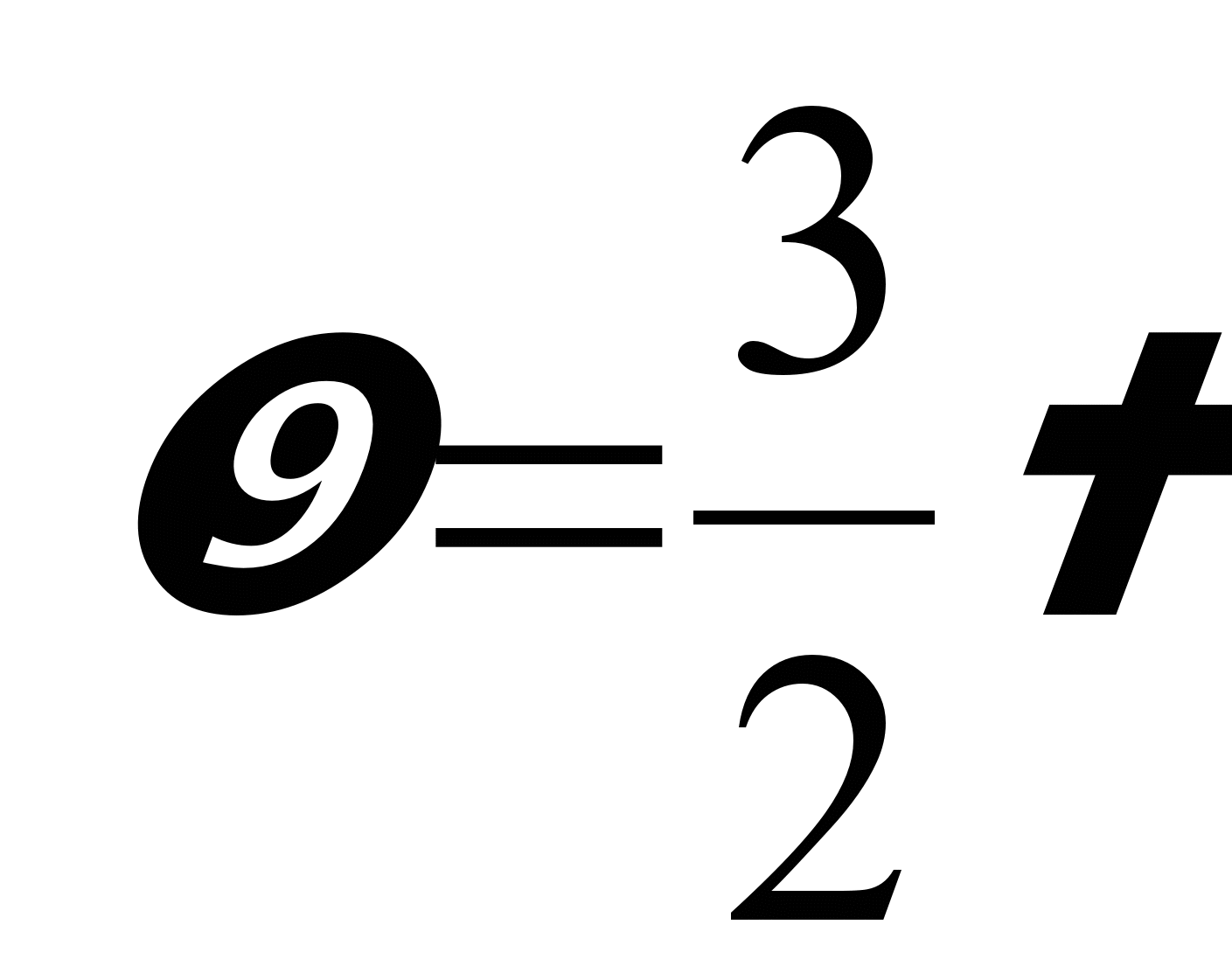
1. Для визуального просмотра корреляционного поля выполните команду **Graphs/Scatter-Dot** (*Графики/Точечное рассеивание*), далее укажите **Simple Scatter**(*Простое рассеивание*), щелкните по кнопке **[Define]** (*Определить*), в область оси X перенесите переменную**Затраты**, а в область оси Y – переменную **Количество туристов**. Нажмите ОК.
2. Сохраните в своей папке полученный отчет под именем *Анализ\_Пирсон.spo.*

Формула (1) для вычисления коэффициента корреляции Пирсона дает достаточно точные результаты и для распределений, не являющихся нормальными и в случае, когда одна из переменных является дискретной. Однако, при таких распределениях, предпочтительнее использовать ранговые коэффициенты корреляции Спирмена или Кенделла.   
  
^

**2. Непараметрические методы анализа.**

Рассмотренный выше метод корреляционного анализа является обоснованным лишь в условиях нормального или близкого к нормальному распределению признаков в изучаемой совокупности. Как видно из формулы (1) для определения линейного коэффициента корреляции необходимо знать значения факторного *X*и результативного *Y*признаков.   
  
В некоторых случаях можно встретиться с такими качествами, которые не поддаются выражению числом единиц. В этом случае прибегают к непараметрическим методам, позволяющим измерить интенсивность связи как между количественными признаками, форма распределения которых отличается от нормальной, так и между качественными признаками.  
  
В основу непараметрических методов положен принцип нумерации значений статистического ряда. Каждой единице совокупности присваивается порядковый номер в ряду, который будет упорядочен по уровню признака. Таким образом, ряд значений ранжируется, а номер каждой отдельной единицы будет ее рангом.  
  
Можно получить предварительное представление о наличии связи между признаками, если сопоставить последовательность взаимного расположения рангов факторного и результативного признаков. Для этого ранги индивидуальных значений факторного признака располагают в порядке возрастания, и если ранги результативного признака обнаруживают тенденцию к увеличению, можно предполагать наличие прямой связи. Если же с увеличением рангов факторного признака ранги результативного признака уменьшаются, то это говорит о возможном наличии между изучаемыми признаками обратной связи.  
  
Для проведения непараметрического анализа используют:

1. Коэффициент ранговой корреляции Спирмена.
2. Коэффициент ранговой корреляции Кенделла.
3. Коэффициент конкордации ω.

**^ 1. Коэффициент ранговой корреляции Спирмена**является непараметрическим аналогом коэффициента корреляции Пирсона и основан на рассмотрении разности рангов значений факторного и результативного признаков. Формула коэффициента ранговой корреляции Спирмена имеет вид:  
 (3)  
где ,  - разности между рангами переменных *X* и *Y*;  
  
*n* – количество признаков.  
  
**^ 2. Коэффициент ранговой корреляции Кенделла** также является мерой связи между переменными *X* и *Y.*Формула коэффициента ранговой корреляции Спирмена имеет вид:  
  
 (4)  
  
где *S =P Q*  
  
Для вычисления  нужно упорядочить ряд рангов переменной X, приведя его к ряду натуральных чисел. Затем рассматривают последовательность рангов переменной Y.  
  
**Первое слагаемое Р** – это мера соответствия последовательности рангов переменной *^ Y* последовательности рангов переменной *X*. При определении слагаемого **P** надо установить, сколько числе, находящихся справа от каждого из элементов последовательности рангов переменной Y, имеет величину ранга, превышающую ранг рассматриваемого элемента.  
  
**Второе слагаемое Q –**это меранесоответствия последовательности рангов переменной *^ Y* последовательности рангов переменной *X.* Для определения Q подсчитывают, сколько чисел, находящихся справа от каждого из членов последовательности рангов переменной Y имею ранг меньше, чем эта единица. Такие величины берутся со знаком минус.  
  
При достаточно большом числе наблюдений между коэффициентами корреляции Спирмена и Кенделла существует соотношение  (5).  
^

**Задание 2. Использование непараметрических методов анализа.**

**Условие задачи:***Эксперты аналитического центра оценивали шансы кандидатов в депутаты на этапе предвыборной кампании следующим образом:*

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| № **депутата** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| **^ Ранг кандидатов по результатам оценки экспертов** | 7 | 4 | 1 | 3 | 10 | 5 | 9 | 2 | 8 | 6 |
| **^ Ранг депутата по числу поданных голосов на выборах** | 5 | 6 | 2 | 7 | 8 | 3 | 10 | 1 | 9 | 4 |

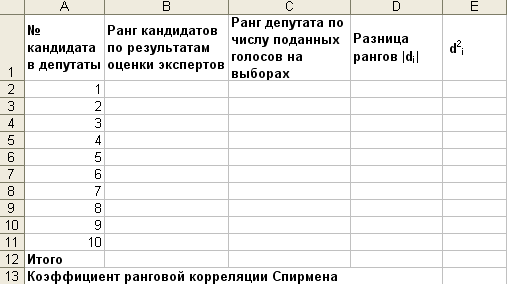
*Необходимо:*

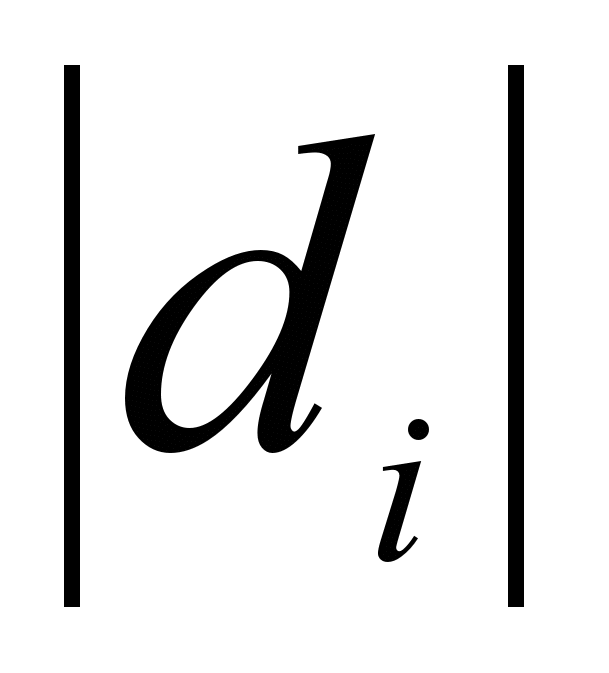
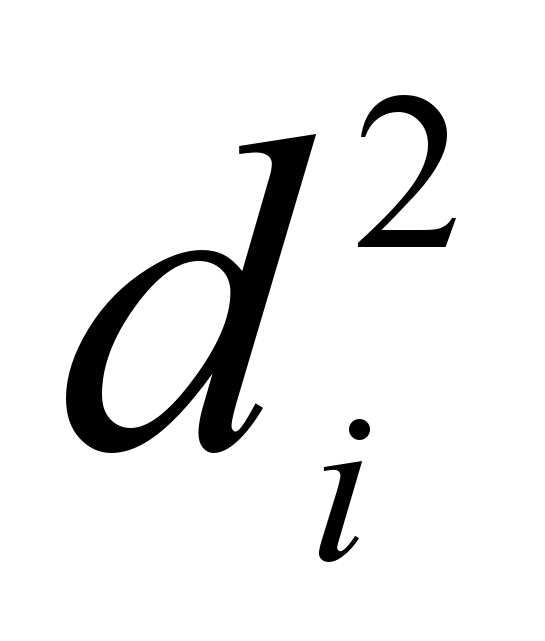
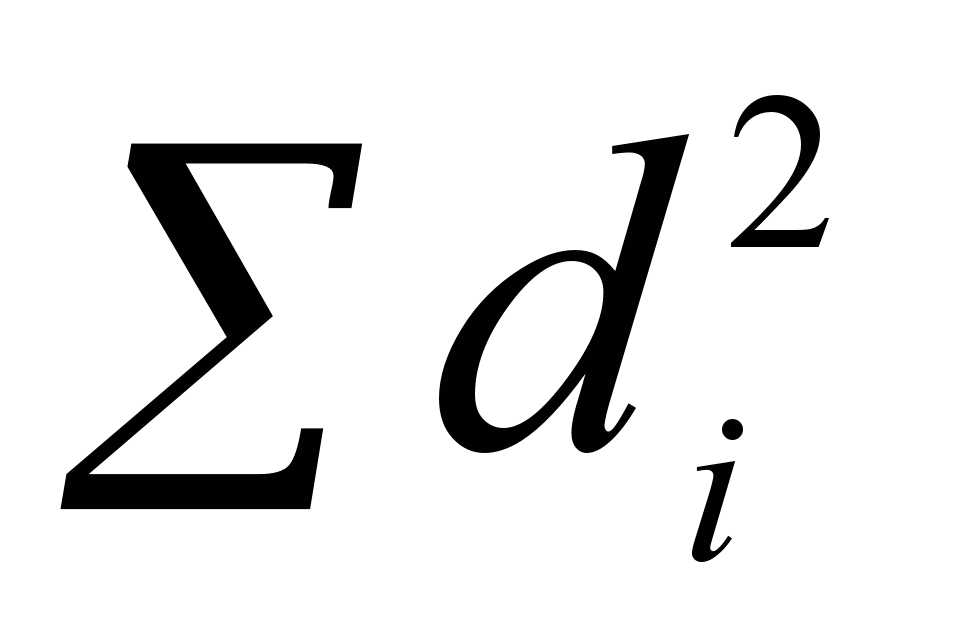
* 1. *Рассчитать коэффициент ранговой корреляции Спирмена по формуле (3) в MS Excel.*
  2. *Рассчитать коэффициент ранговой корреляции Кенделла по формуле (4) в MS Excel.*
  3. *Провести непараметрический анализ в ППП SPSS.*

^

**2.1 Расчет коэффициента корреляции Спирмена с использованием формулы (3).**

1. Откройте новую книгу MS Excel и создайте таблицу согласно рис. 9.
2. **Лист 1** переименуйте в **Спирмен** и сохраните в своей папке под именем *Непараметр\_анализ.xls*.

  
  
**Рис.9. Данные для задачи №2 (пункт 2.1)**

1. В диапазоне ячеек **D2:D11** рассчитайте абсолютную разницу рангов  с использованием математической функции **ABS.**
2. В диапазоне ячеек **E2:E11** рассчитайте квадрат разницы рангов .
3. В ячейке **Е12** рассчитайте .
4. В ячейке **Е13** рассчитайте коэффициент ранговой корреляции Спирмена по формуле (3).

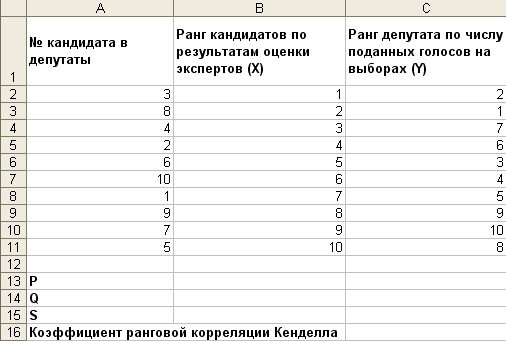
http://www.studmed.ru/docs/static/7/5/6/3/a/7563a69eb8f.png

1. Сделайте вывод о направлении и тесноте связи между исследуемыми признаками.

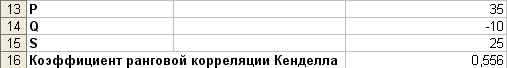
^

**2.2 Расчет коэффициента корреляции Кенделла с использованием формулы (4).**

1. Перейдите на новый лист книги *Непараметр\_анализ.xls.*и дайте ему имя **Кенделл**.
2. Скопируйте на этот лист данные из диапазона ячеек A1:C11 листа **Спирмен.**
3. Выделите данный диапазон ячеек и проведите его сортировку по **столбцу В** (ранжирование по переменной *X*). Дополните лист данными согласно рис. 10.

  
  
**Рис.10. Данные для задачи №2 (пункт 2.2).**

1. Рассчитайте слагаемое **Р – меру соответствия** последовательности рангов переменной *Y* последовательности рангов переменной *X.* Например, первому значению в последовательности рангов переменной Y, т.е. числу 2 соответствует восемь чисел (7, 6, 3, 4, 5, 9, 10, 8), которые превышаю ранг 2; второму значению 1 соответствует также восемь чисел (7, 6, 3, 4, 5, 9, 10, 8), превышающих 1; третьему значению 7 соответствует три числа (9, 10, 8) и т.д. Суммируя таким образом полученные числа (8 8 3 ……) получите значение **Р** в ячейке **С13**(см. рис.11).
2. Рассчитайте слагаемое **Q – меру несоответствия** последовательности рангов переменной *^ Y* последовательности рангов переменной *X*. Например, первому значению 2 последовательности рангов переменной Y, соответствует только одно число 1, которое меньше ранга 2; второму значению 1 соответствует 0 чисел, ранг которых меньше 1; третьему значению 7 соответствуют четыре числа (6, 3, 4, 5), ранг которых меньше 7 и т.д. Суммируя таким образом полученные числа со знаком минус (-1 – 0 – 4 -…), получите значение **Р** в ячейке **С14**(см. рис.11).
3. В ячейке **С15** рассчитайте значение **S**(см. рис.11).
4. В ячейке **С16** рассчитайте значение коэффициента ранговой корреляции Кенделла (см. рис.11).

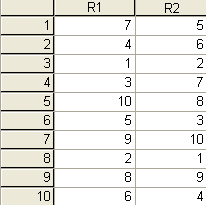
  
  
**Рис.11. Расчет коэффициента ранговой корреляции Кенделла.**

1. Сравните коэффициенты Спирмена и Кенделла, используя формулу (5).

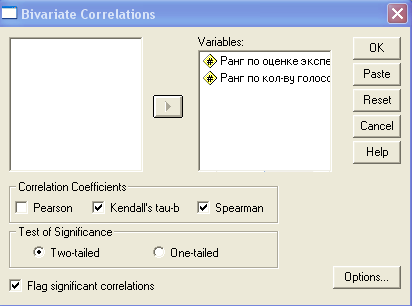
В большинстве случаев рекомендуется применять коэффициент корреляции Спирмена. Использование коэффициента Кенделла оправдано только в том случае, когда в структуре данных имеются выбросы.  
  
^

**2.3 Расчет коэффициента корреляции Спирмена и Кенделла в ППП SPSS.**

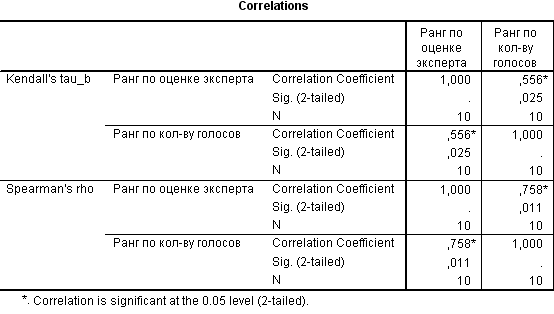
1. Загрузите SPSS и введите данные в режиме **Data View***(Просмотр данных)* согласно рис. 12. Сохраните в своей папке под именем *Непараметрический\_Анализ. Sav (*данные можно также скопировать из файла MS Excel).
2. Перейдите в режим **Variable View***(Просмотр переменных)*и для переменных R1 и R2 задайте **^ Label***(Метка)* **Ранг по оценке эксперта** и **Ранг по полученным голосам** соответственно.

  
  
**Рис.12. Данные в режиме** *Data View.*

1. Выполните команду **Analyze/Correlate/Bivariate** *(Анализ/Корреляция/Парные).*
2. В диалоговом окне в область переменных перетащите обе переменные, укажите коэффициент корреляции Спирмена и Кенделла, уровень значимости – двусторонний (см**.**рис. 13).

  
  
**Рис.13. Диалоговое окно «Корреляции».**

1. Проанализируйте полученный результат (рис. 14) аналогично Заданию №1. Сравните полученные результаты с коэффициентами корреляции Спирмена и Кенделла, рассчитанными по формулам в MS Excel.
2. Сохраните отчет в своей папке под именем *Непараметр*\_*Анализ.spo.*

  
  
**Рис.14. Отчет по корреляции Спирмена и Кендалла.**  
  
^

**Задание для самостоятельной работы.**

Установить, зависит ли количество посетителей музея и посетителей парка от числа ясных дней за определенный период. Для этого:

* Вычислить коэффициенты корреляции с использованием функции КОРРЕЛ в MS Excel.
* Провести исследование в ППП SPSS.
* Построить корреляционное поле.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **^ Число ясных дней (Х)** | 8 | 14 | 20 | 25 | 20 | 15 |
| **^ Количество посетителей музея (Y)** | 495 | 503 | 380 | 305 | 348 | 465 |
| **^ Количество посетителей парка (Y)** | 132 | 348 | 643 | 865 | 743 | 541 |

^

**3. Регрессионный анализ.**

Цель регрессионного анализа – определить количественные связи между зависимыми случайными величинами. Одна из этих величин полагается зависимой и называется **откликом,** другие – независимые, называются **факторами**или**регрессорами**.   
  
Регрессионный анализ позволяет:

* Производить расчет регрессионных моделей путем определения значений параметров – постоянных коэффициентов при независимых переменных-регрессорах.
* Проверять гипотезу об адекватности модели имеющимся наблюдениям.
* Использовать модель для прогнозирования значений зависимой переменной при новых или ненаблюдаемых значениях независимых переменных.

Среди регрессионных моделей обычно выделяют:

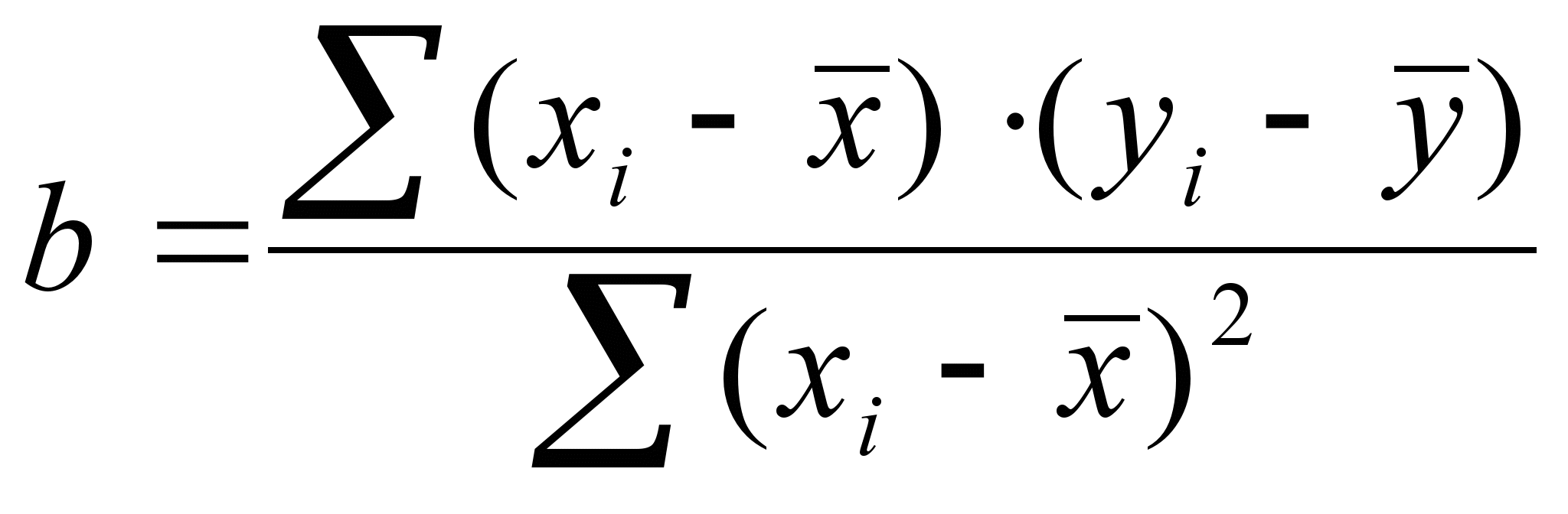
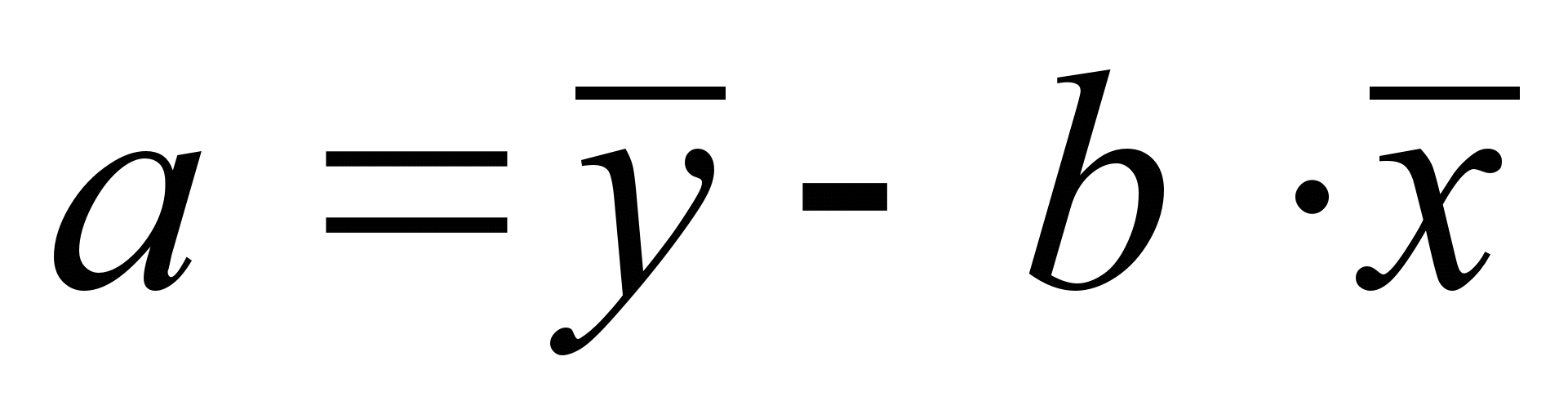
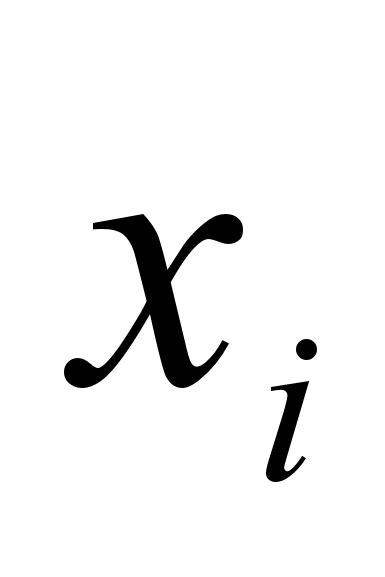
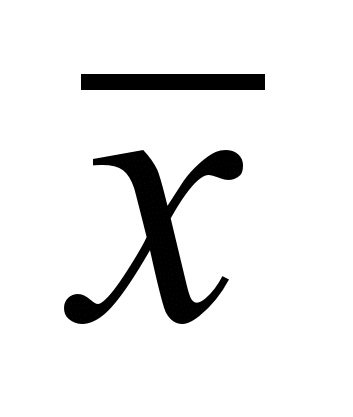
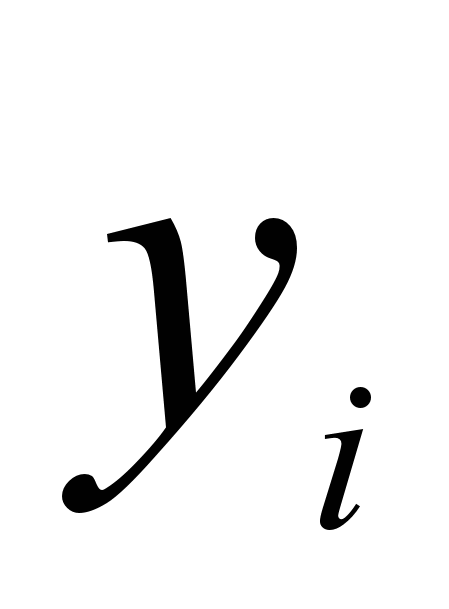
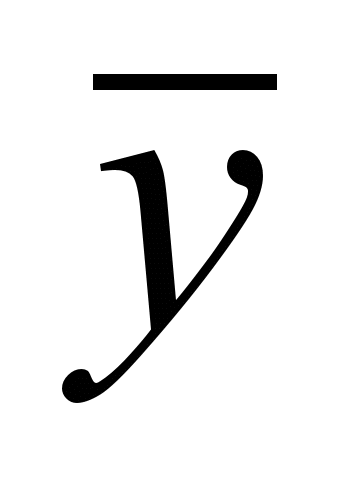
1. Однопараметрические модели (зависимости от одной переменной).
2. Многопараметрические модели (зависимости от нескольких переменных).
3. Линейные модели относительно независимых переменных.
4. Модели нелинейные по переменным и нелинейные по параметрам.

^

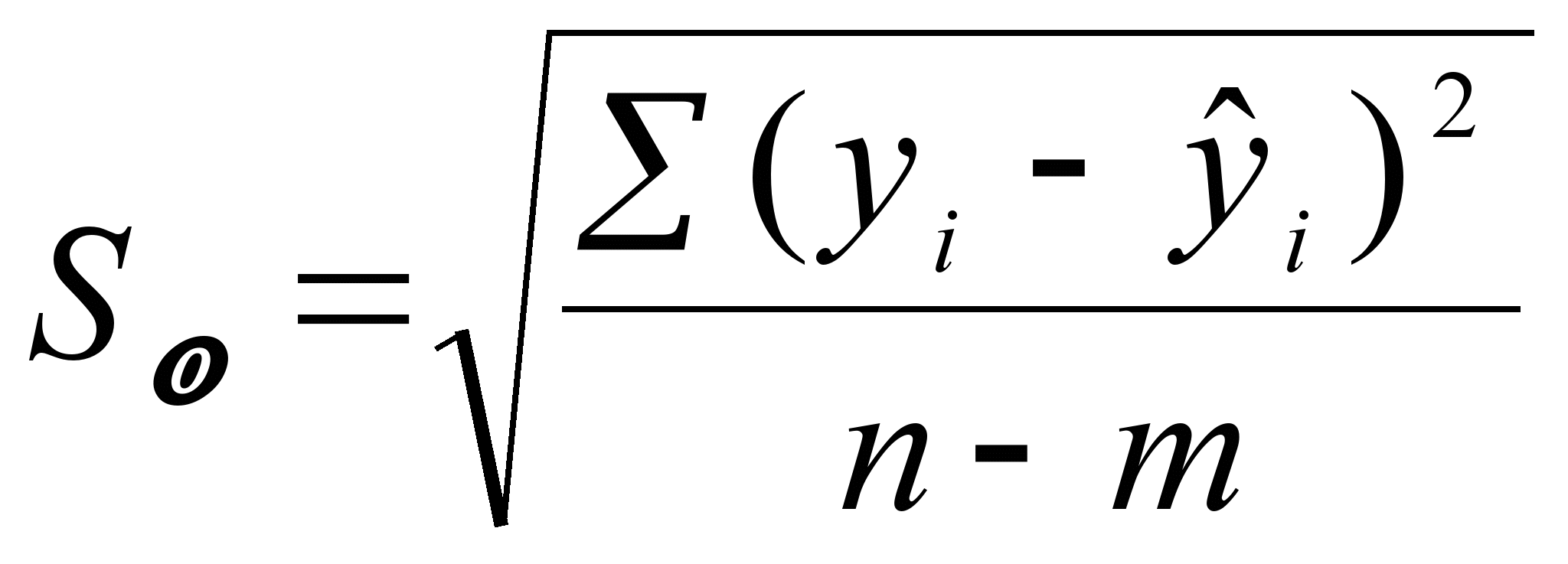
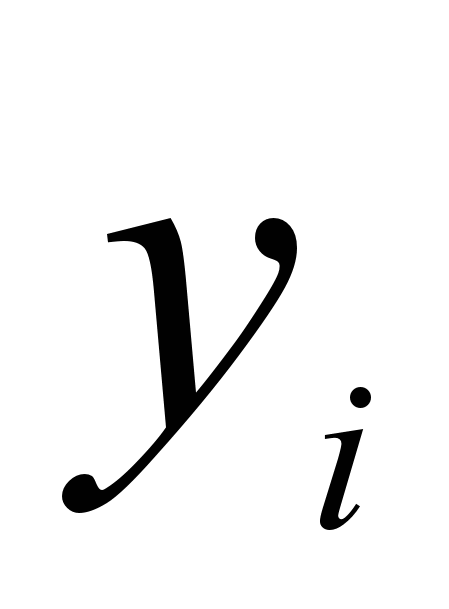
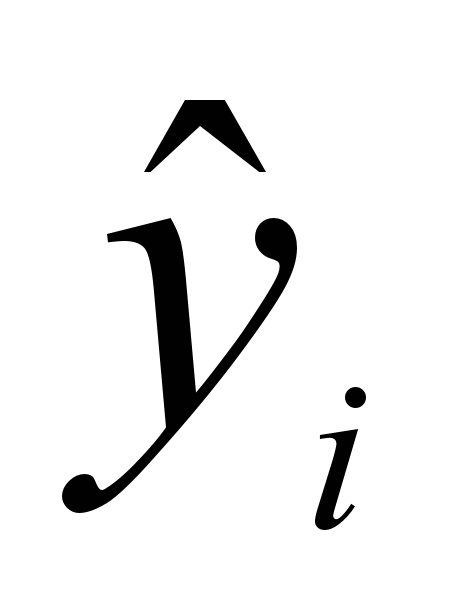
**3.1 Однопараметрические модели.**

Наиболее распространенные методы регрессионного анализа являются параметрическими, большая их часть основана на предположении о нормальном распределении данных наблюдений, поэтому в каждом случае анализа необходима предварительная проверка данных нормальному распределению.  
  
**^ Линейная модель.**  
  
Если коэффициент корреляции по абсолютной величине близок к единице, то для построения зависимости используется линейная модель. Для других случаев используются более сложные нелинейные модели.  
  
**^ Модель простой линейной регрессии**имеет вид:  
  
*y=b\*x a*(6)  
  
где *y* – зависимая переменная (отклик),  
  
*х –*независимая переменная-регрессор (фактор),  
  
*b* – коэффициент регрессии,  
  
*a* – смещение по оси ординат.   
  
Смещение по оси ординат соответствует точке на оси *^ Y* (вертикальной оси), где прямая регрессии пересекает эту ось. Коэффициент регрессии *b* через соотношение *b = tg(a)* указывает на угол наклона прямой.   
  
Для получения параметров модели *а* и *b*чаще всего используют метод наименьших квадратов, основанный на минимизации среднеквадратической ошибки модели. Т.е. оптимальным решением задачи построения линейной регрессии является такая прямая, для которой сумма квадратов вертикальных расстояний до отдельных точек данных является минимальной.   
  
При выполнении регрессионного анализа необходимо получить оценки, позволяющие дать прогноз с определенной точностью и вероятностью. При нормальном законе распределения условия будут удовлетворены, если оценить:

1. **Ожидаемые значения коэффициентов *b*и *а.***

 (7)  (8)  
  
где  - значения независимого признака;  
  
 - среднее значение независимого признака;  
  
 - фактические значения результативного признака, полученные по данным наблюдений;  
  
 - среднее фактических значений результативного признака.

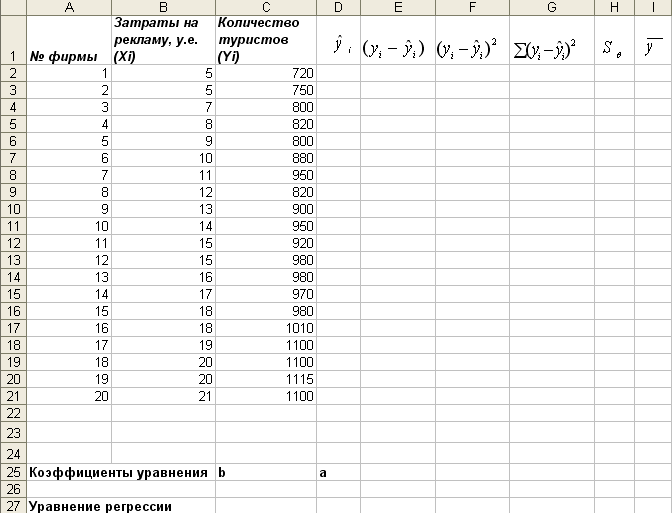
1. **Стандартную (среднюю квадратическую) ошибку модели.**

 (9),   
  
где  - фактические значения результативного признака, полученные по данным наблюдений;  
  
 - рассчитанные значения результативного признака;  
  
*n -*объем выборки;  
  
*m –*число параметров в уравнении регрессии.  
^

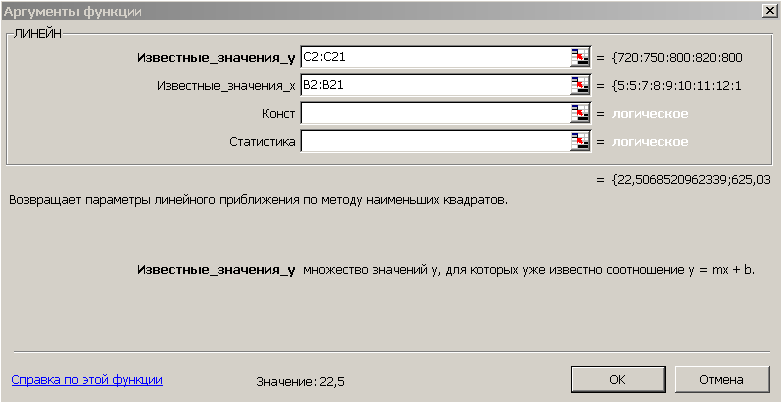
**Задание 3. Построение модели линейной регрессии.**

**Условие задачи:***По 20 туристическим фирмам были установлены затраты на рекламную кампанию и количество туристов, воспользовавшихся после ее проведения услугами каждой фирмы. Необходимо построить модель линейной регрессии, учитывая, что переменные подчинены нормальному закону распределения.*  
  
^

**3.1 Расчет параметров уравнения линейной регрессии с использованием функции ЛИНЕЙН.**

Откройте новую книгу MS Excel и создайте таблицу согласно рис. 15, сохраните в своей папке под именем *Регрессионный\_Анализ.xls*. (Данные можно также скопировать из файла по заданию №1).  
  
  
  
**Рис.15. Данные для задачи №3.**

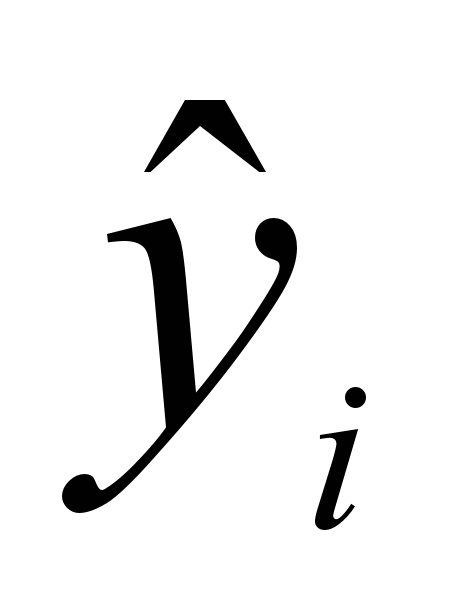
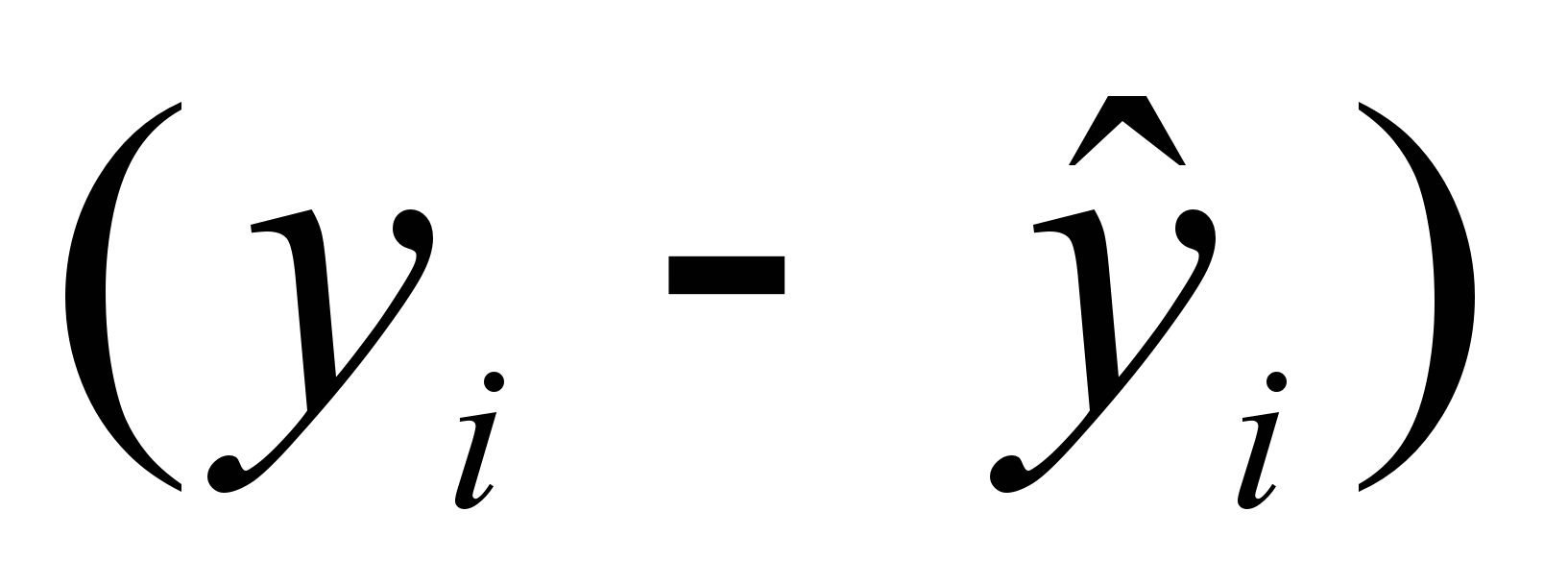
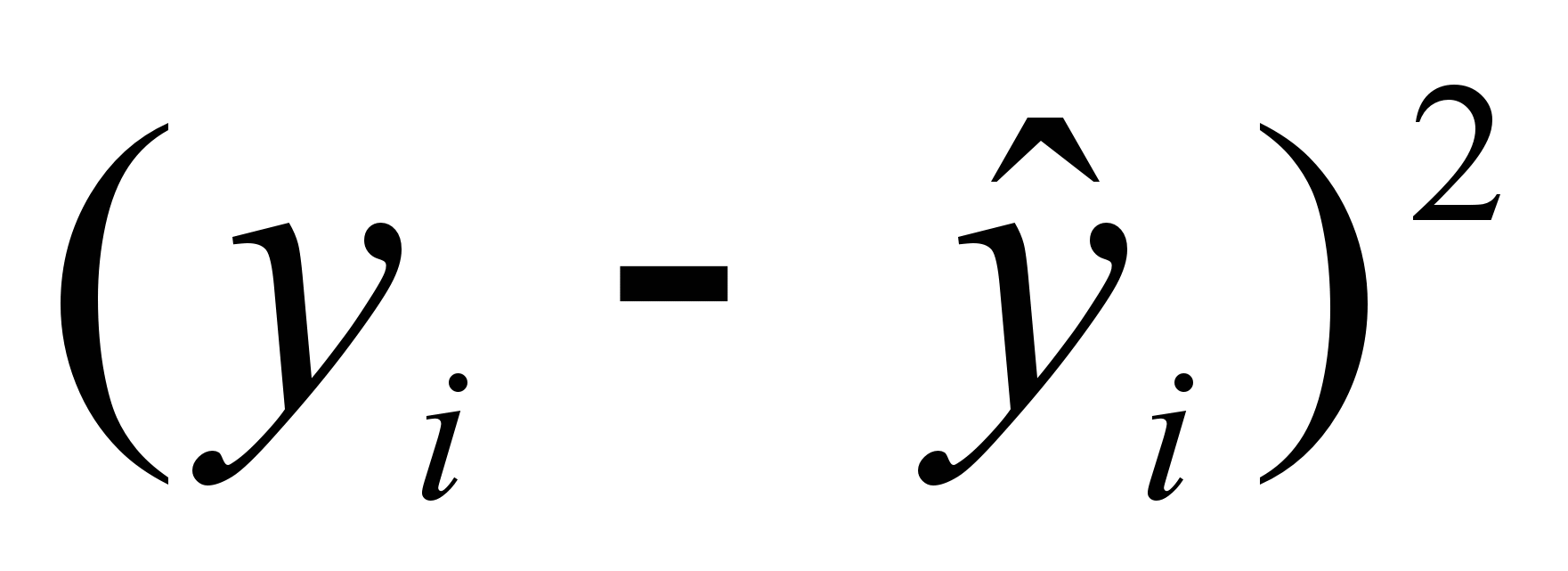
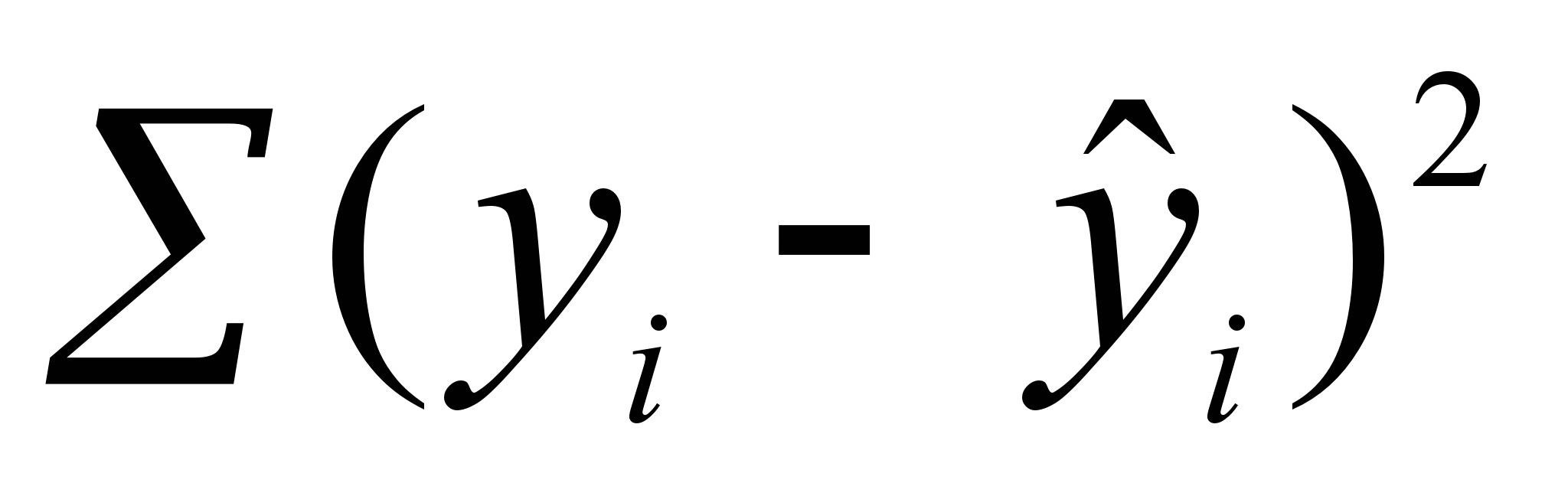
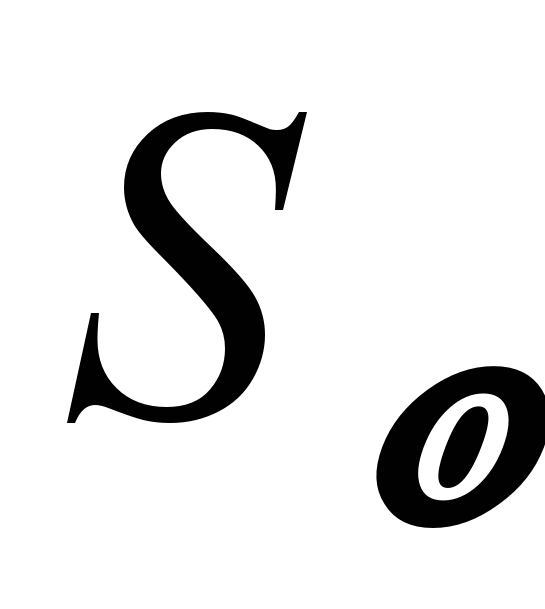
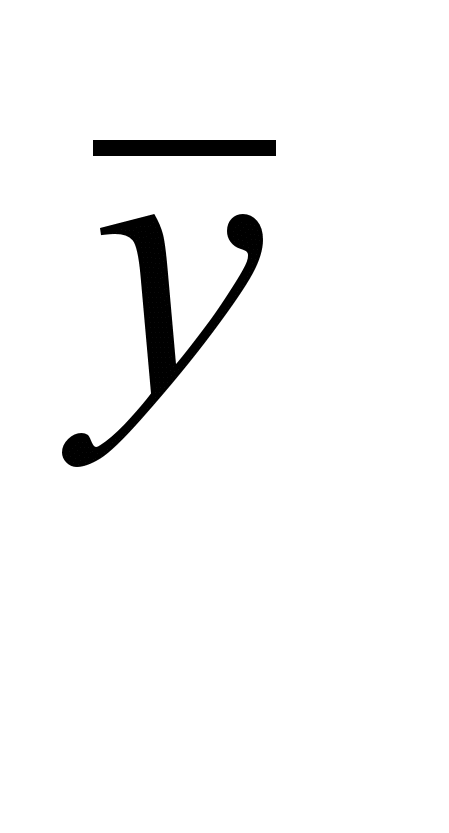
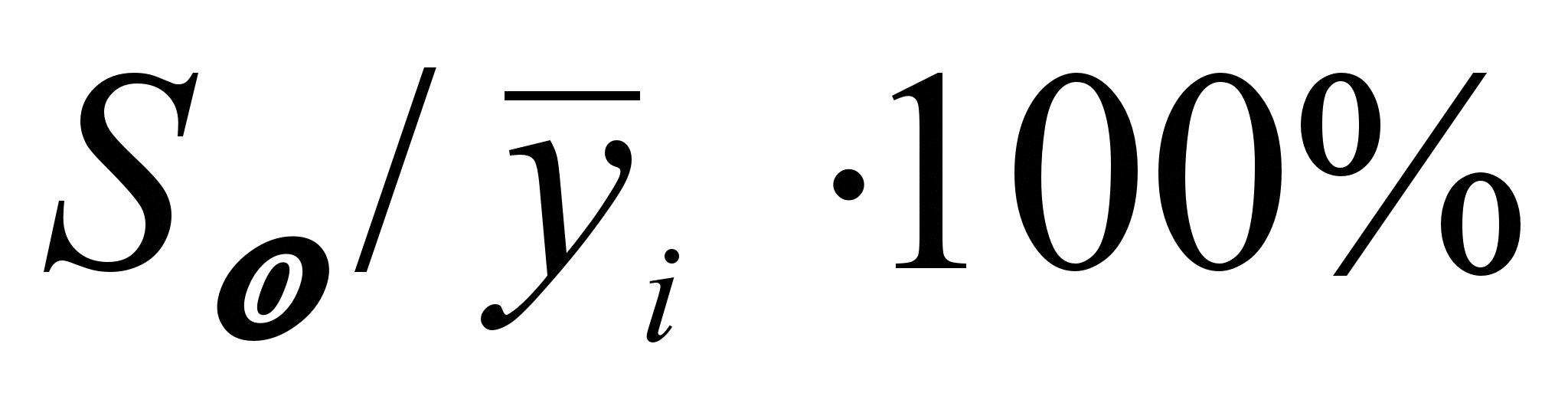
1. Для получения коэффициентов **а** и **b** линейного уравнения регрессии***y=b\*x a***, описывающего зависимость количества привлеченных туристов от затрат на рекламу воспользуемся статистической функцией **ЛИНЕЙН.** Для этого выделите две ячейки **C26:D26** и выполните вставку функции **ЛИНЕЙН**с аргументами согласно рис.16. Здесь **Известные\_значения\_y** – диапазон значений **Количество туристов**, **Известные\_значения\_x** – диапазон значений **Затраты на рекламу.** Нажмите комбинацию клавиш SHIFT CTRL ENTER.

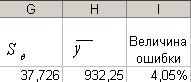
  
  
**Рис. 16. Аргументы функции ЛИНЕЙН.**

1. В ячейку **C27** введите уравнение регрессии ***y=b\*x a***, (вместо *b* и *a*подставьте полученные коэффициенты линейной регрессии).

http://www.studmed.ru/docs/static/2/e/a/4/e/2ea4ee00bb8.png

1. Рассчитайте стандартную (среднюю квадратическую) ошибку модели по формуле (7). Для этого:

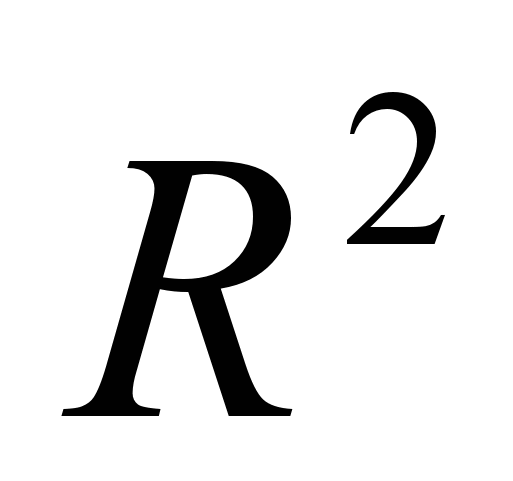
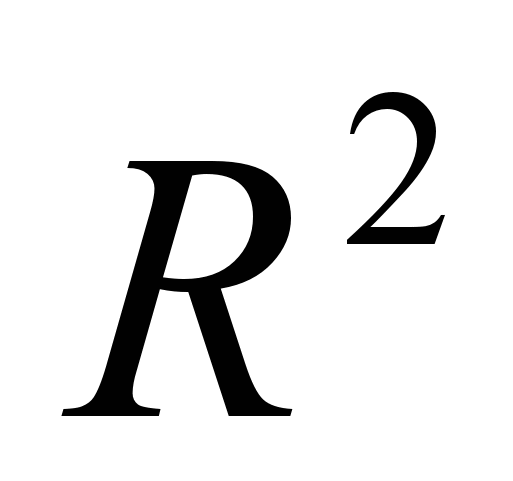
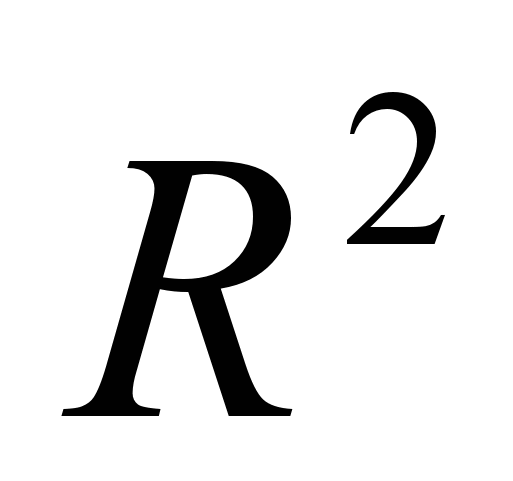
* В диапазоне ячеек **D2:D21** рассчитайте значения результативного признака  путем подстановки значений независимого признака-регрессора *Х* в уравнение линейной регрессии.
* В диапазоне ячеек **E2:E21** рассчитайте отклонения фактических значений результативного признака от рассчитанных значений .
* В диапазоне ячеек **F2:F21** рассчитайте квадраты отклонений .
* В ячейке **G2** рассчитайте сумму квадратов отклонений .
* В ячейке **Н2** рассчитайте стандартную ошибку модели  по формуле 9.
* Проанализируйте величину ошибки. Для этого рассчитайте среднее значение  фактического результативного признака в ячейке **I2** и найденное значение подставьте в формулу .

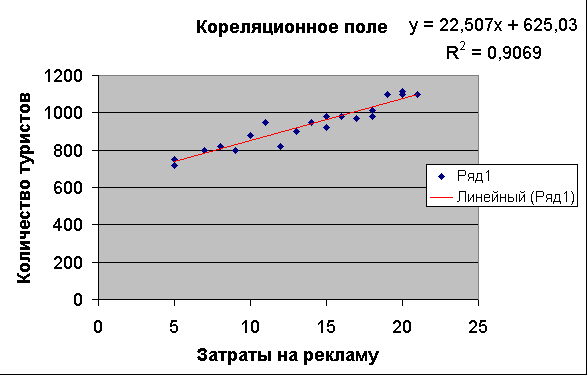


1. Сделайте вывод об адекватности линейной модели.

^

**3.2 Нахождение уравнения линейной регрессии графическим методом.**

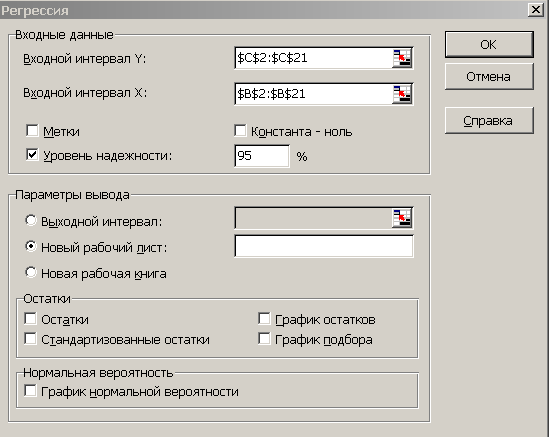
1. Для получения уравнения регрессии построим **корреляционное поле** переменных **X** (затраты на рекламу) и **Y** (количество туристов).
2. Выделите диапазон ячеек **В2:С21,** запустите мастера диаграмм и выберите тип диаграммы – **Точечная.**Задайте для диаграммы имя – *Корреляционное поле*, ось Х – *Затраты на рекламу*, ось Y – *Количество туристов.* На последнем шаге мастера укажите место расположения – *отдельный лист.*
3. Добавьте **линию тренда** на точечный график. Для этого необходимо выделить диаграмму и выполнить команду меню **^ Диаграмма /Добавить линию тренда,**либо выполнить данную команду из контекстного меню, щелкнув по любой точке графика. **Линия тренда**– графическое представление направления изменения ряда данных
4. Выберите тип тренда **Линейный,**который используется для аппроксимации данных по методу наименьших квадратов в соответствии с уравнением ***y=b\*x a***.
5. На вкладке **Параметры** установите флажки **Показать уравнение на диаграмме**и **Поместить на диаграмму величину достоверности аппроксимации .**Щелкните по кнопке ОК. **-**это число от 0 до 1, которое отражает близость линии тренда к фактическим данным. Линия тренда наиболее соответствует действительности, когда значение ** близко к 1.**
6. Сравните уравнение регрессии, полученное графическим методом (рис. 17), с уравнением, рассчитанным с помощью функции ЛИНЕЙН.

  
  
**Рис.17. Модель линейной регрессии.**  
  
^

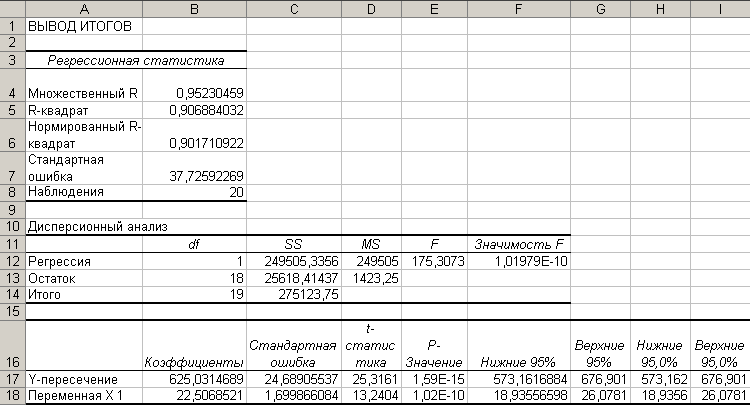
**3.3 Построение модели линейной регрессии с помощью инструмента «Регрессия».**

1. Сначала убедитесь, что был активизирован **Пакет анализа,**т.е. в меню **Сервис** есть команда **Анализ данных**. Если нет, то выполните команду **Сервис/Надстройки.**В диалоговом окне **Надстройки**установите флажок **Пакет анализа** и щелкните по кнопке ОК.
2. Далее выполните команду **Сервис/Анализ данных.** Выберите инструмент анализа **Регрессия** из списка **Инструменты анализа**. Щелкните по кнопке ОК.
3. На экране появится диалоговое окно **Регрессия** (рис.18).

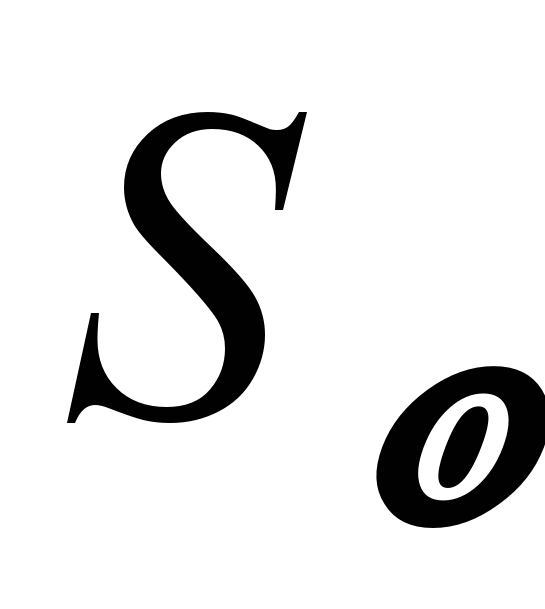
* в текстовом поле **^ ВХОДНОЙ ИНТЕРВАЛ Y** введите диапазон со значениями зависимой переменной **$C$2:$C$21.**
* в текстовом поле **ВХОДНОЙ ИНТЕРВАЛ Х** введите диапазон со значениями независимых переменных **$В$2:$В$21.**
* Убедитесь, что в поле **Уровень надежности**введено **95 %** и переключатель **Параметры вывода** установлен в положении **Новый рабочий лист**.
* Щелкните по кнопке **ОК**.

  
  
**Рис. 18. Диалоговое окно инструмента анализа Регрессия.**

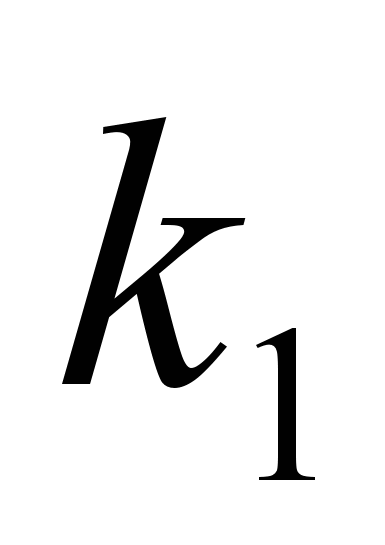
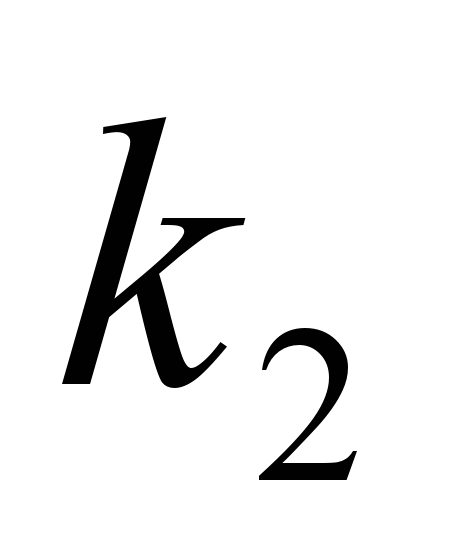
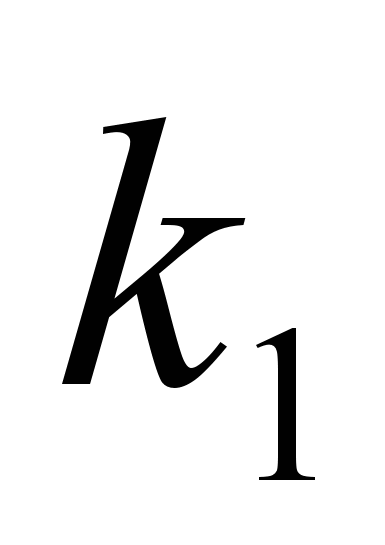
1. В результате на новом листе будет отображены результаты использования инструмента **Регрессия** (рис.19).

  
  
**Рис.19. Вывод итогов инструмента Регрессия.**

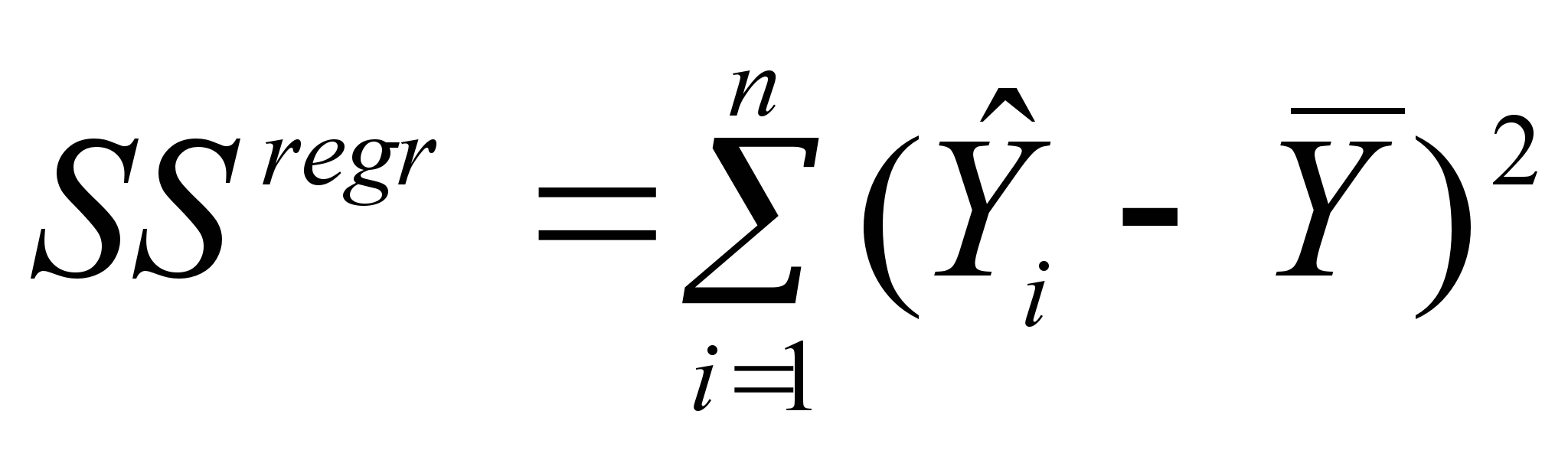
1. В области **Регрессионная статистика** получили:
   * **Множественный R=0,952** (коэффициент корреляции по Пирсону) Сравните этот результат с коэффициентом корреляции, полученным в задании №1.

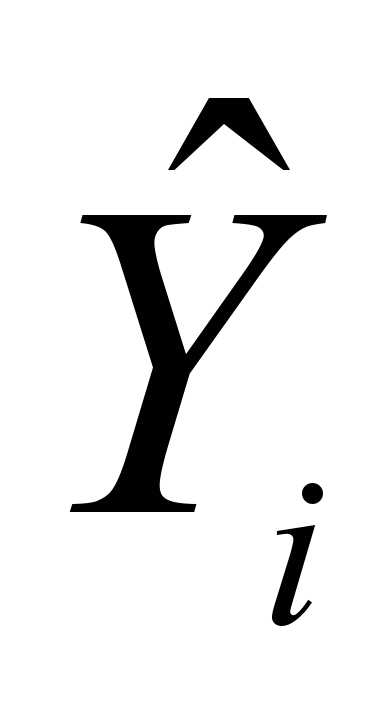
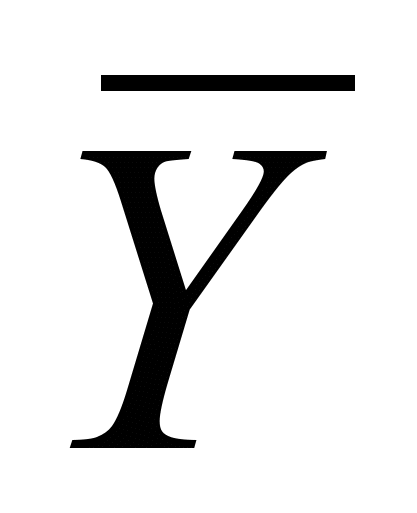
* **R-квадрат = 0,906.**Эта величинаносит название **коэффициента детерминации** – это частное от суммы квадратов, обусловленных регрессией и остаточной суммы квадратов. Эта величина характеризует качество регрессионной прямой, то есть степень соответствия между регрессионной моделью и исходными данными. Мера определённости всегда лежит в диапазоне от 0 до 1. В нашем случае R2= 0,907 – это значит, что регрессионной моделью описано 90,7 % случаев. В простом линейном регрессионном анализе квадратный корень из коэффициента детерминации, равен коэффициенту корреляции Пирсона. *Сравните полученную величину с величиной достоверности аппроксимации, полученной при построении модели линейной регрессии графическим способом.*
* **Стандартная ошибка = 37,73 –**оценка адекватности линейной модели. Сравните этот результат с величиной , полученной в задании №4 (по формуле (9)).

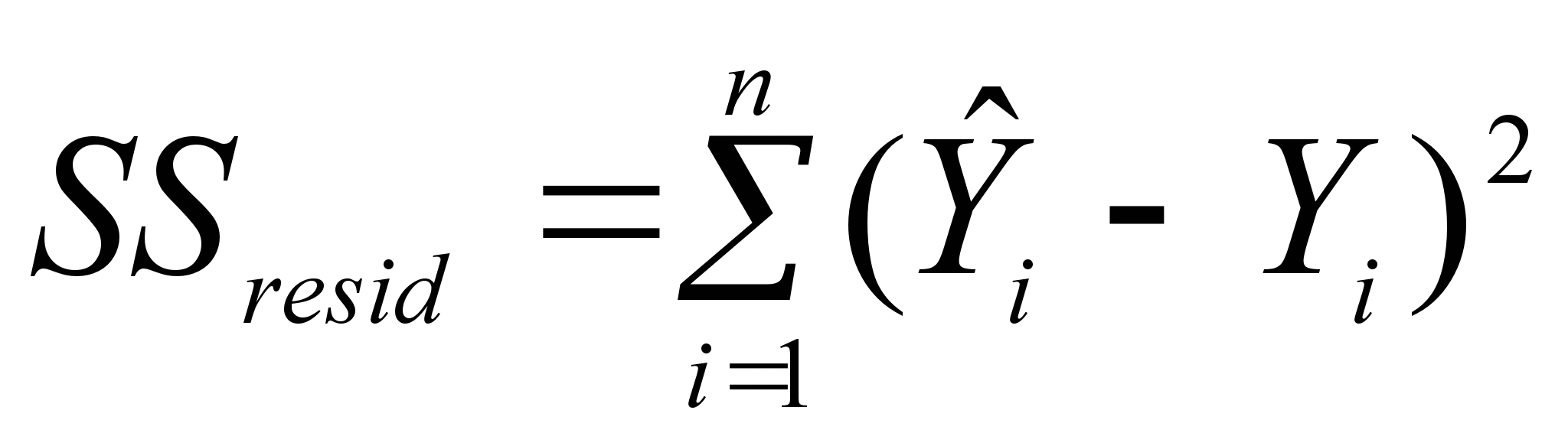
   
В области **Дисперсионный анализ** мы получили:   
  
6.1 Столбец**df -** число степеней свободы (используется при проверке адекватности модели по статистическим таблицам):

* в строке **Регрессия**находится – количество коэффициентов уравнения, не считая свободного члена b;
* в строке **Остаток**находится =n--1, где n – количество исходных данных.

6.2. Столбец**SS (сумма квадратов):**

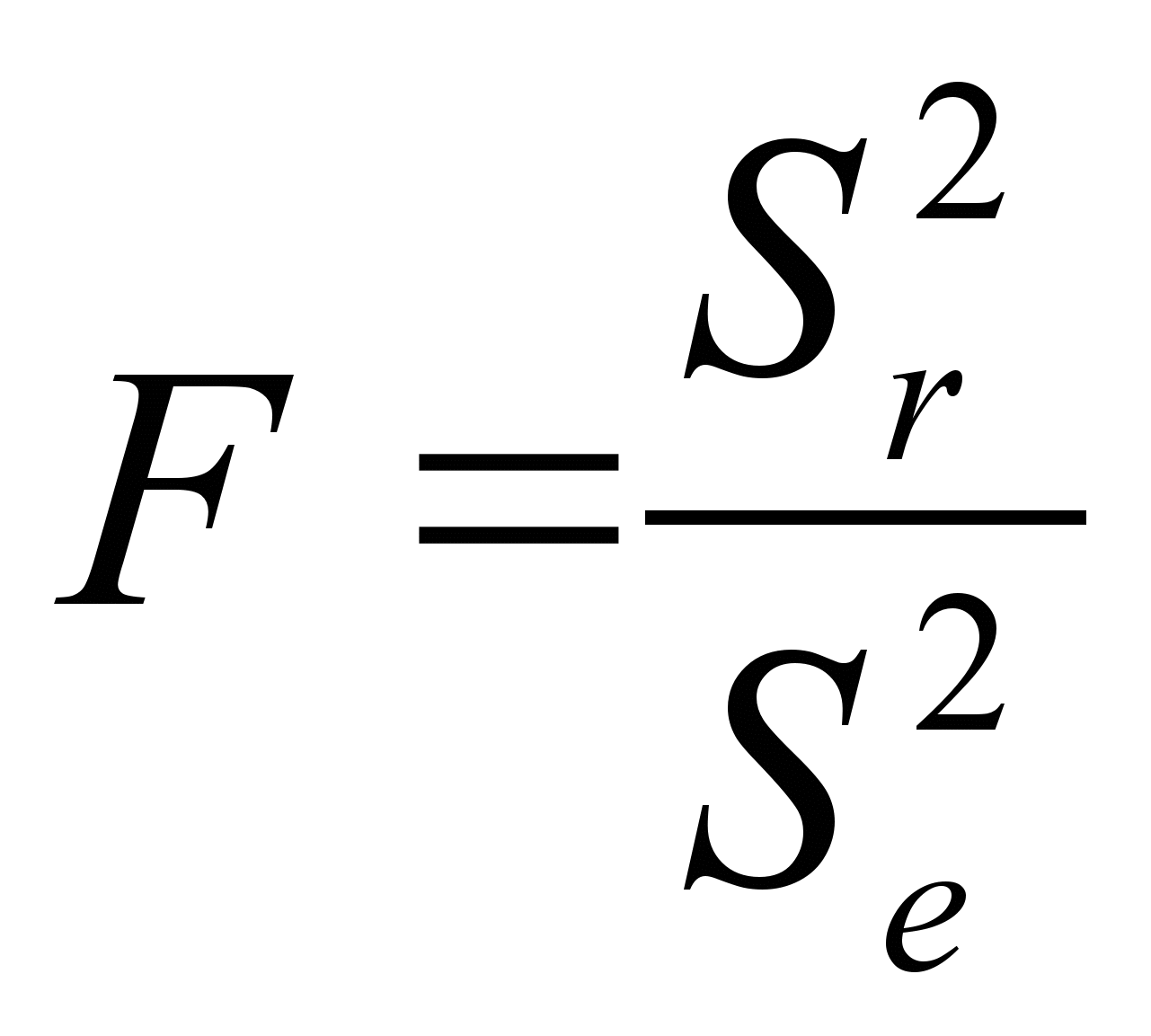
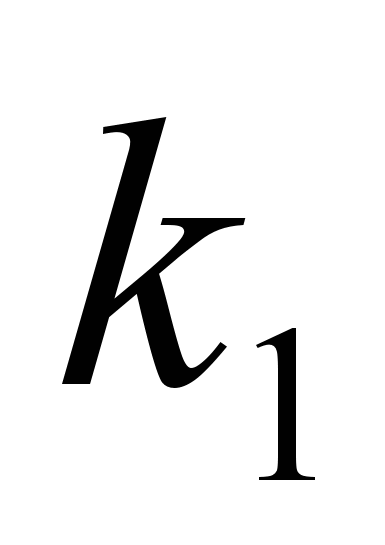
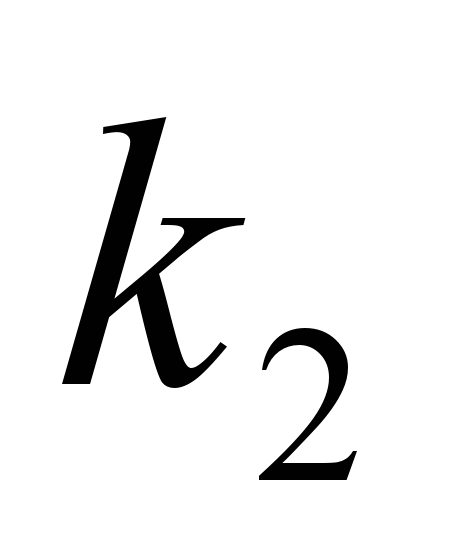
* в строке **Регрессия:**  - приведена доля дисперсии, которая описывается уравнением регрессии (сумма квадратов, обусловленная регрессией).

где  - модельные (расчетные) значения Y, полученные путем подстановки значений Х в построенную модель;  
  
- среднее значение Y;

* в строке **Остаток: ** - приведена доля дисперсии, которая не учитывается при записи уравнения (остаточная сумма квадратов).

6.3. Столбец**MS - вспомогательные величины:**

* в строке Регрессия: ;
* в строке Остаток: .

Используются для расчета критерия Фишера.  
  
6.4. Столбец**F - критерий Фишера**.   
.  
  
6.5. Столбец **Значимость F (критерия Фишера)**- **оценка адекватности построенной модели**. Находится по значениям F,  и  с помощью функции **FРАСП.** Если Значимость **F** меньше 0,05, то модель может считаться адекватной с вероятностью 0,95.

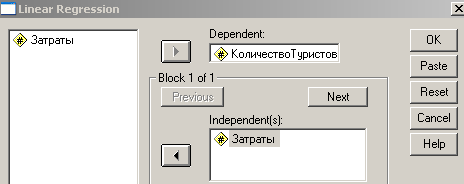
1. Среди полученных результатов есть столбец **«Коэффициенты»,** содержащий значение **a = 625,031** в строке **«Y-пересечение», b=22,507** – в строке **«Переменная Х1»**. *Сравните полученные результаты с ранее рассчитанными коэффициентами****a****и****b.***
2. **Стандартная ошибка**, **t-статистика** - это вспомогательные величины, используемые для проверки значимости коэффициентов модели.
3. **^ Р - величина** - оценка значимости коэффициентов модели. Если Р - величина меньше 0,05, то с вероятностью 0,95 можно считать, что соответствующий коэффициент модели значим (т.е. его нельзя считать равным нулю и Y значимо зависит от соответствующего Х).
4. - доверительные интервалы для коэффициентов модели.%Нижние и верхние 95

^

**3.4 Построение модели линейной регрессии с помощью ППП SPSS.**

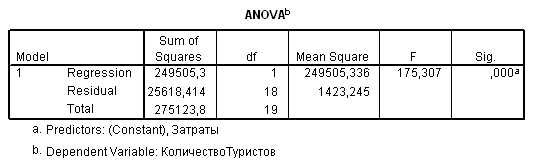
Для построения модели линейной регрессии в SPSS следуйте следующим инструкциям:

1. Выполните команду **Analyze**/**Regression/Linear** *(Анализ/Регрессия/Линейная).*
2. В диалоговом окне переменную **Количество туристов** перенесите в область **Dependent** *(Зависимая),* а переменную **Затраты** – в область **Independent***(Независимая).*Нажмите кнопку **ОК**(см. рис. 20).

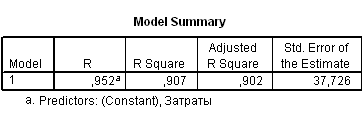
  
  
Рис.20. Фрагмент окна **Linear Regression.**

1. Проанализируем полученные результаты.
   1. Первое, на что необходимо обратить внимание, - это таблица **ANOVA**, на которой представлены результаты дисперсионного анализа (рис. 21). Результаты данной таблицы соответствуют результатам, полученным с использованием инструмента анализа **Регрессия**в MS Excel (таблица *Дисперсионный анализ*).

* В строке **Regression***(Регрессия)* приведена доля дисперсии, которая описывается уравнением регрессии (сумма квадратов, обусловленная регрессией).
* В строке **Residual**(*Остаток*) – доля дисперсии, которая не учитывается при записи уравнения (остаточная сумма квадратов).
* Наибольший интерес представляет **статистическая значимость Sig или значимость F**0,001), и, соответственно, адекватность построенной модели стремится к 100%.≤(вероятность случайности результата), которая должна быть меньше или равна 0,05. Если значимость меньше 0,05, то модель может считаться адекватной с вероятностью 0,95. В нашем случае вероятность того, что получен случайный результат, очень мала (p

  
  
Рис.21. Таблица **^ ANOVA (**Дисперсионный анализ).

* 1. Затем следует рассмотреть таблицу **Model Summary** *(Сводная таблица по модели)*, содержащую важные сведения о построенной модели (рис.22). Результаты данной таблицы соответствуют результатам, полученным с помощью инструмента анализа **Регрессия**(таблица *Регрессионная статистика*) в MS Excel.
* Коэффициент корреляции **^ R**=0,952, что указывает на наличие очень сильной связи между исследуемыми переменными.
* Величина**R Square** (R2)**- коэффициент детерминации.**В нашем случае R2= 0,907 – это значит, что регрессионной моделью описано 90,7 % случаев.
* Более точное значение дает **Adjusted R Square** (*Исправленный квадрат R*).
* Третьим, практически значимым показателем, определяющим качество регрессионной модели, является величина стандартной ошибки расчетов **(Std. Error of the Estimate).** *Сравните данный показатель с полученным ранее в MS Excel.*

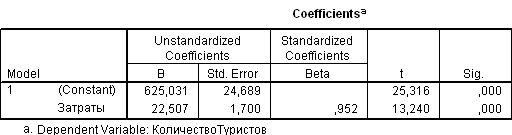
  
  
Рис.22. Таблица **Model Summary**.  
На основании таблиц **ANOVA** и **Model Summary** можно судить о практической пригодности построенной регрессионной модели. Учитывая, что ANOVA показывает весьма высокую значимость (менее 0,001), коэффициент корреляции равен 0,952, можно сделать вывод о том, что с учетом ограничения модель описывает приблизительно 90 % совокупной дисперсии, то есть построенная регрессионная модель является статистически значимой и практически приемлемой.

* 1. Основные практические результаты регрессии содержатся в таблице **Coefficients** (*Коэффициенты*), представленной на рис.23. Результаты данной таблицы соответствуют результатам, полученным с помощью инструмента анализа **Регрессия** в MS Excel.

* Столбец **В** содержит коэффициент регрессии. Он служат для формирования регрессионного уравнения, по которому можно рассчитать величину зависимой переменной при разных значениях независимых. В нашем примере **B=22,507**.
* Cтрока **Constant** (*Постоянная*) содержит значение зависимой переменной при нулевом значении независимой переменной (смещение по оси ординат) и в нашем случае **А = 625,031.**
* Итак, в результате построения регрессионной модели можно сформировать следующее регрессионное уравнение:

**Y=22,507\*X 625,031**

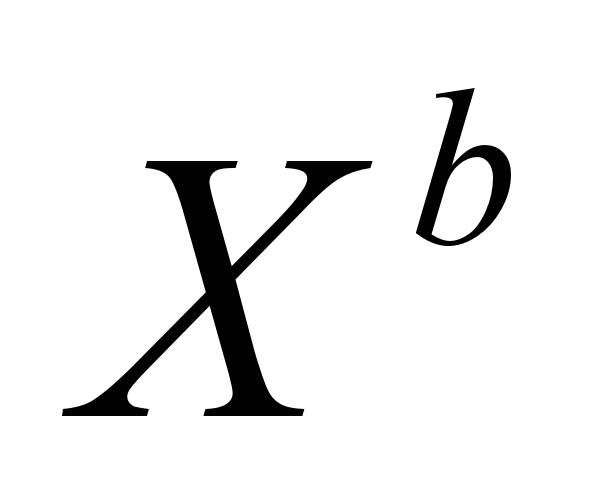
* Столбец **Std. Error** *(Стандартная ошибка),*рассчи­тываемая для коэффициента регрессии **В.** При 95%-ном доверительном уровне коэффициент может отклоняться от величины **В**на ± 2\*Std.Error. Это означает, что **коэффициент В**, равный 22,507 в 95 % случаев может отклоняться от данного значения на ± 2 \* 1,7 или на ± 3,4. Минимальное значение коэффициен­та будет равно 22,507 – 3,4 = 19,107; а максимальное – 22,507 3,4 = 25,907. Таким образом, в 95 % случаев коэффициент регрессии варьируется в пределах от 19,107 до 25,907 (при среднем значении 22,507).

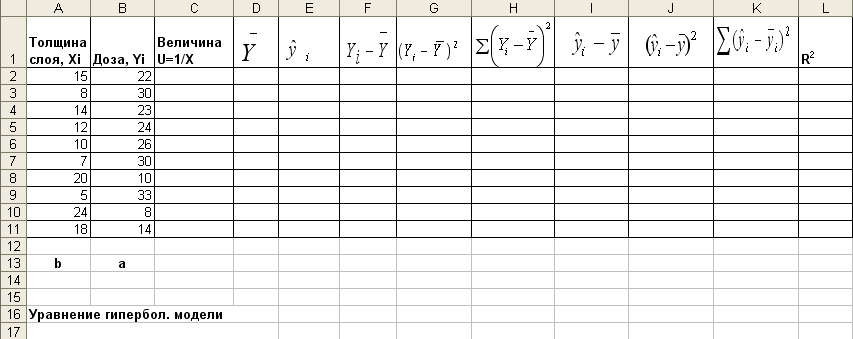
  
  
Рис.23. Таблица **Coefficients**.  
На этом интерпретация результатов регрессионного анализа может считаться за­вершенной.   
^

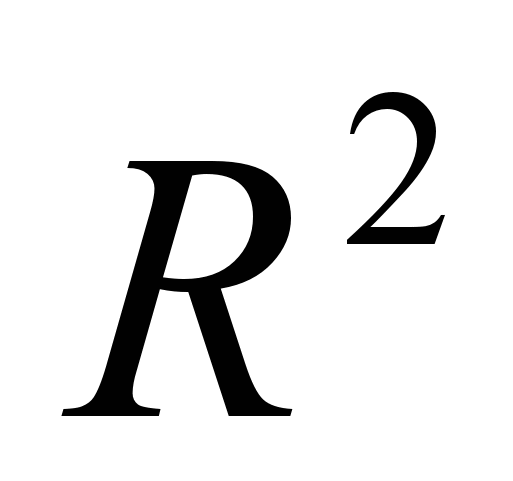
**3.5. Построение линии регрессии.**

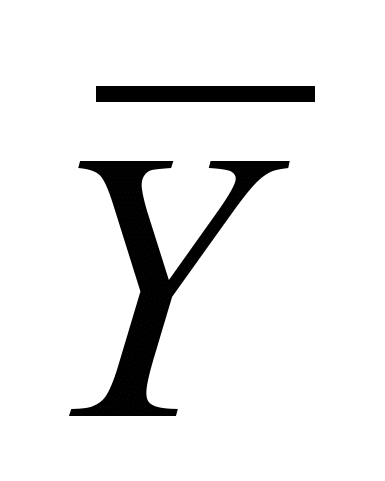
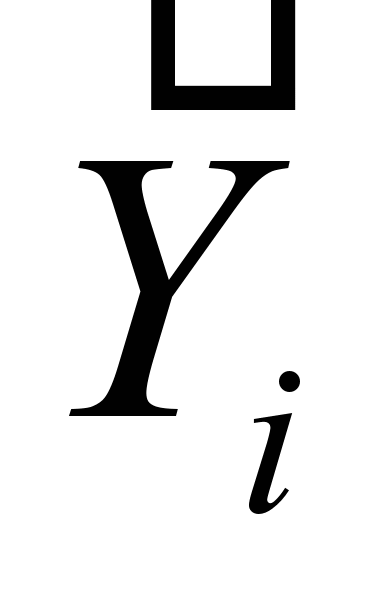
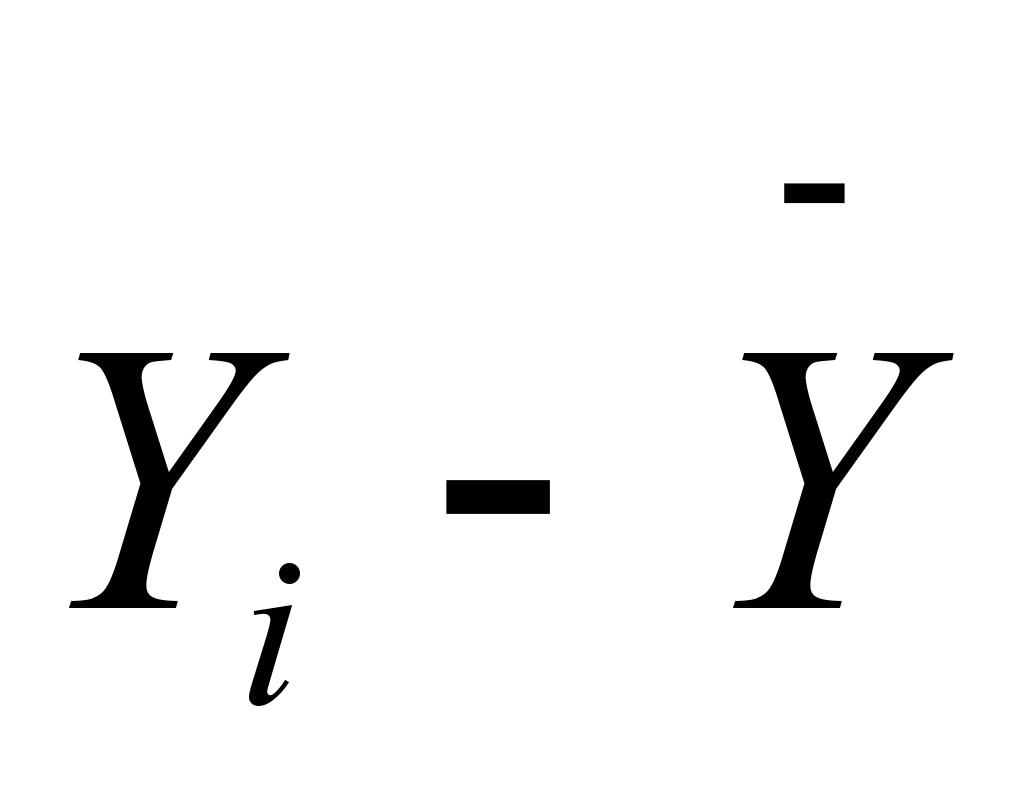
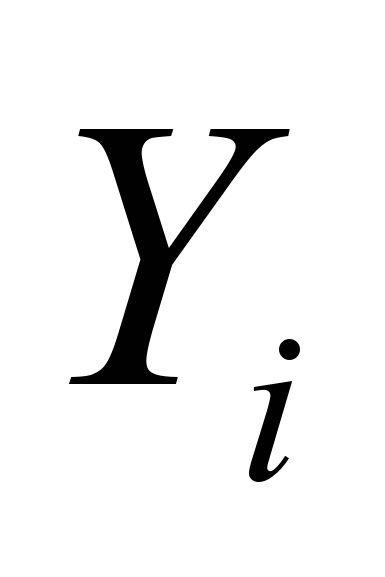
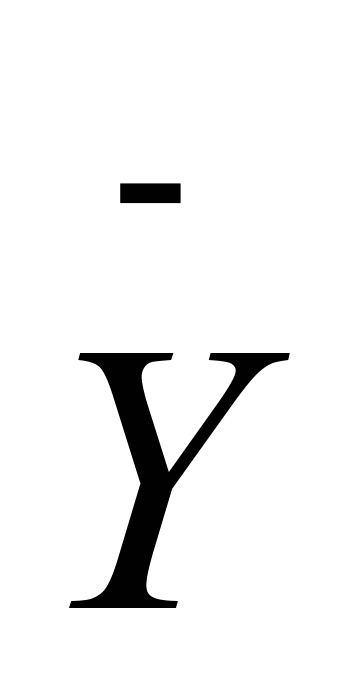
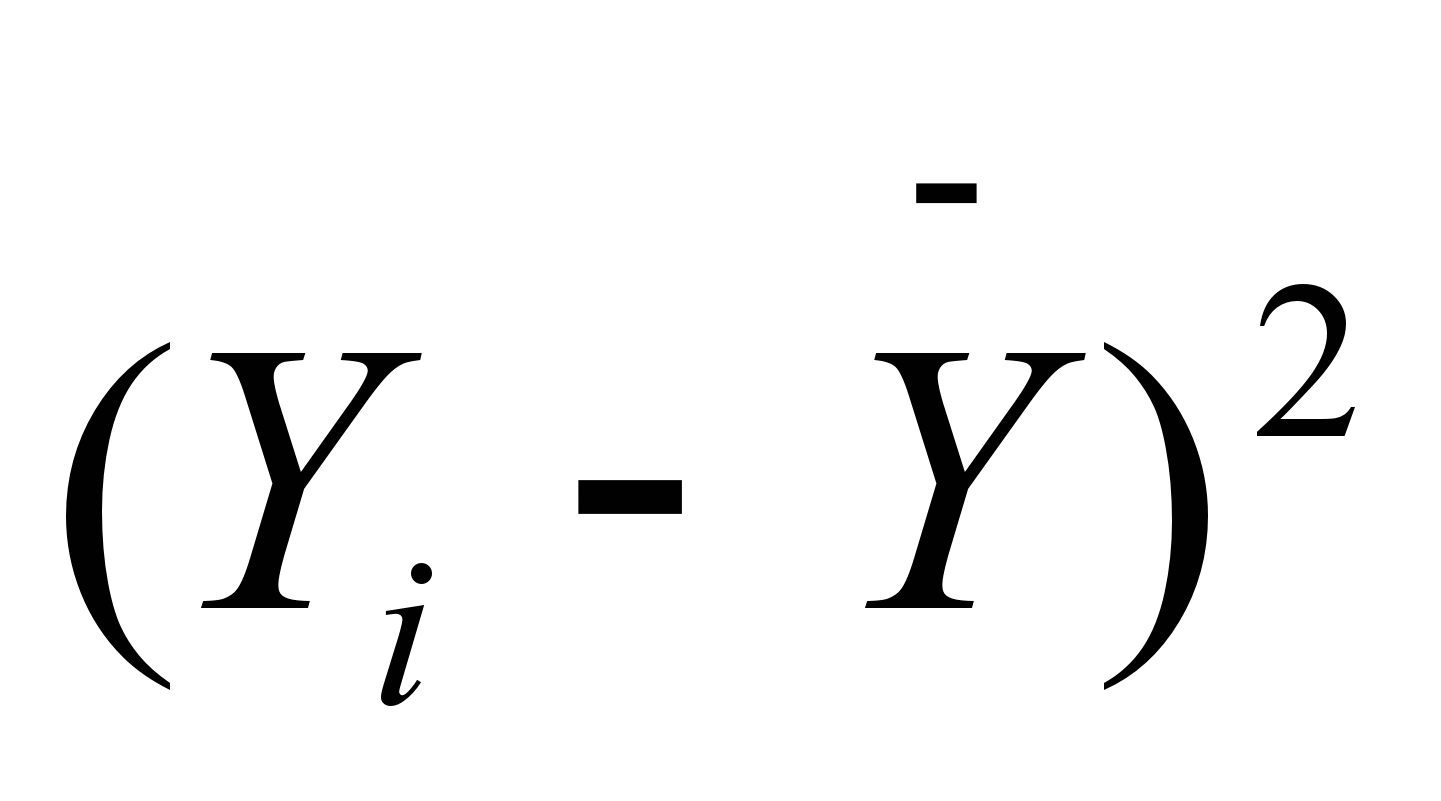
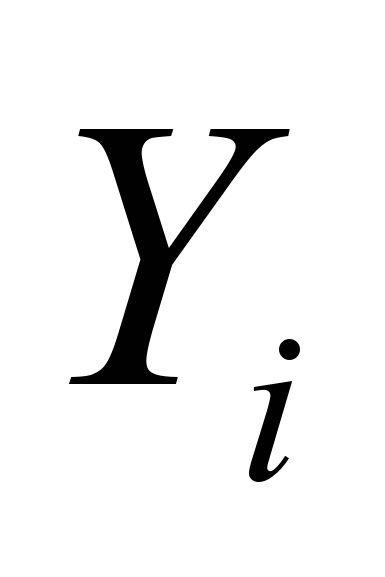
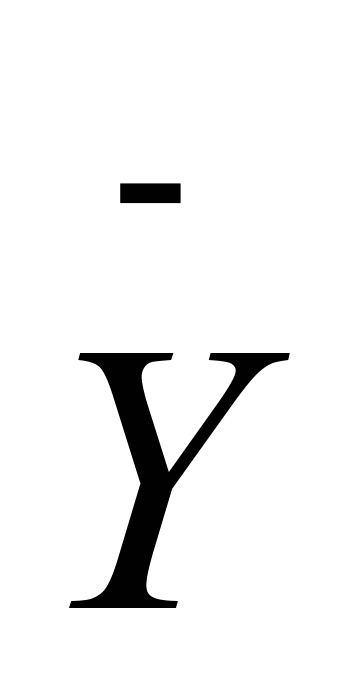
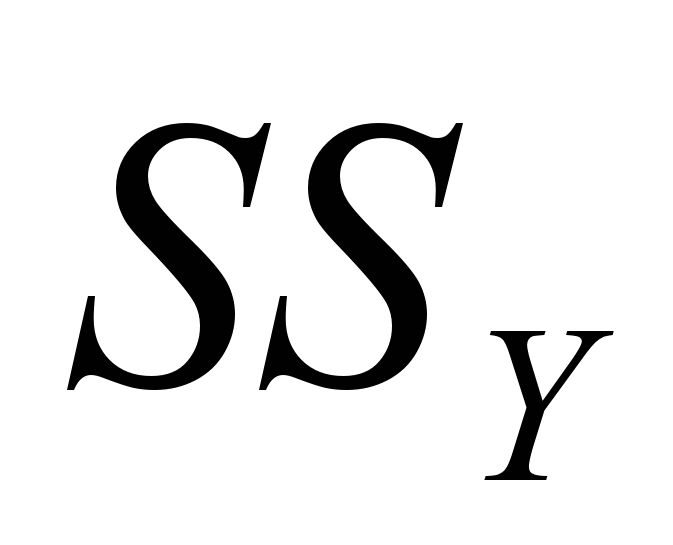
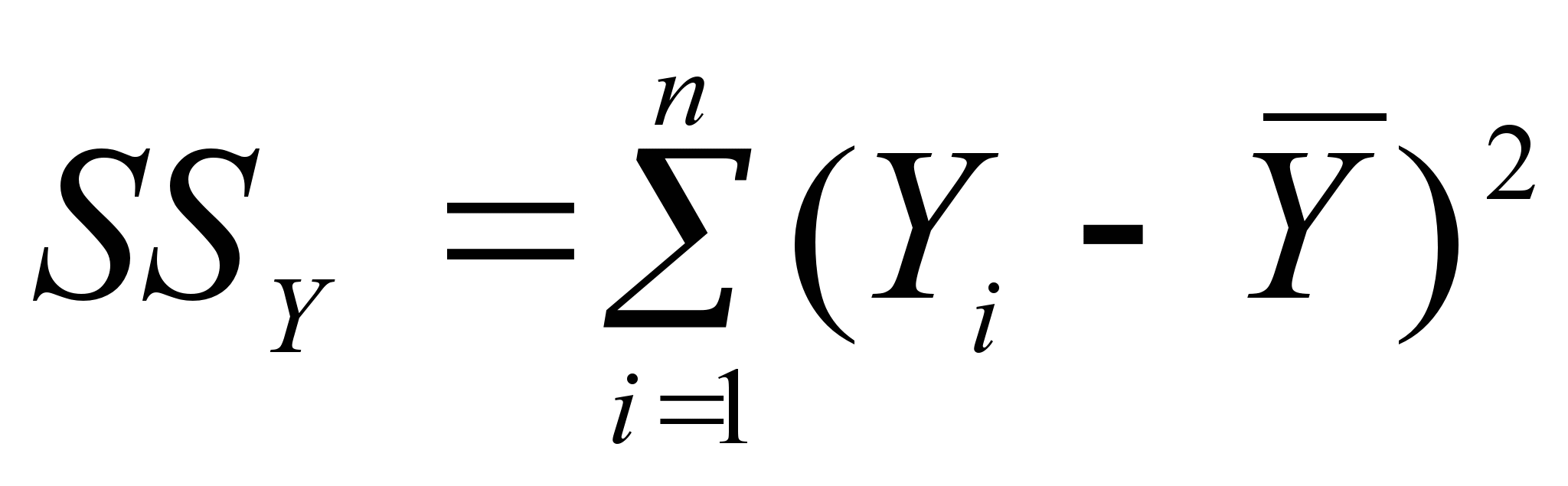
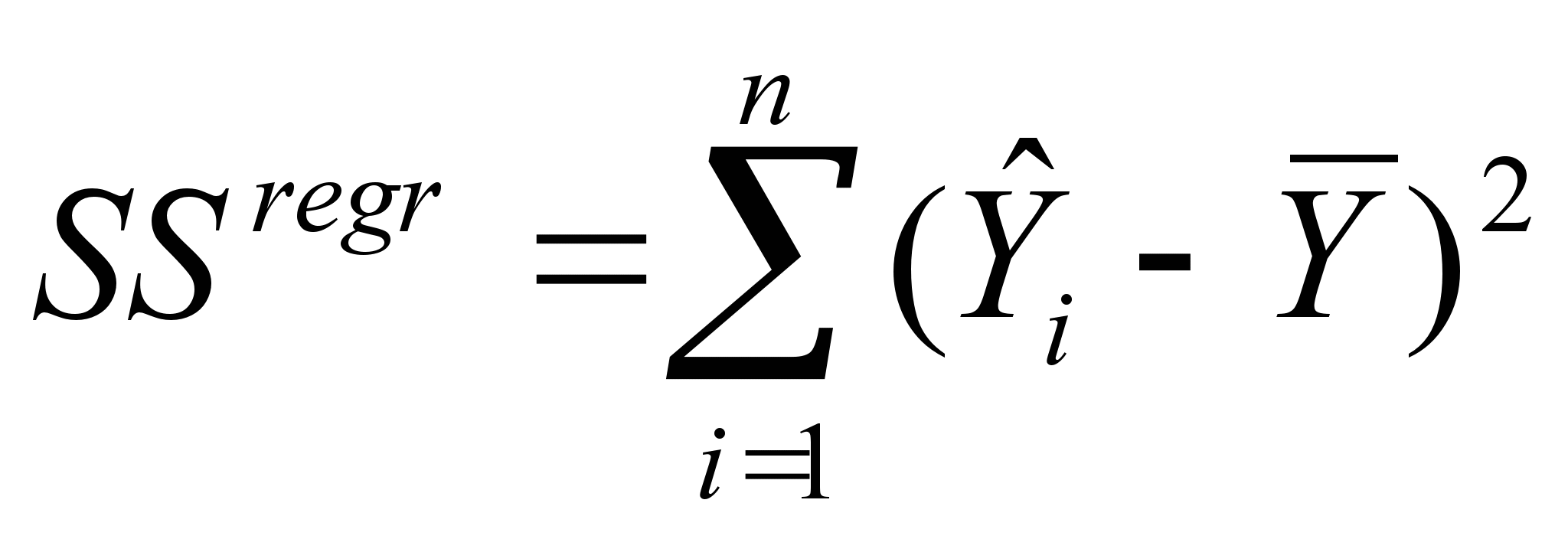
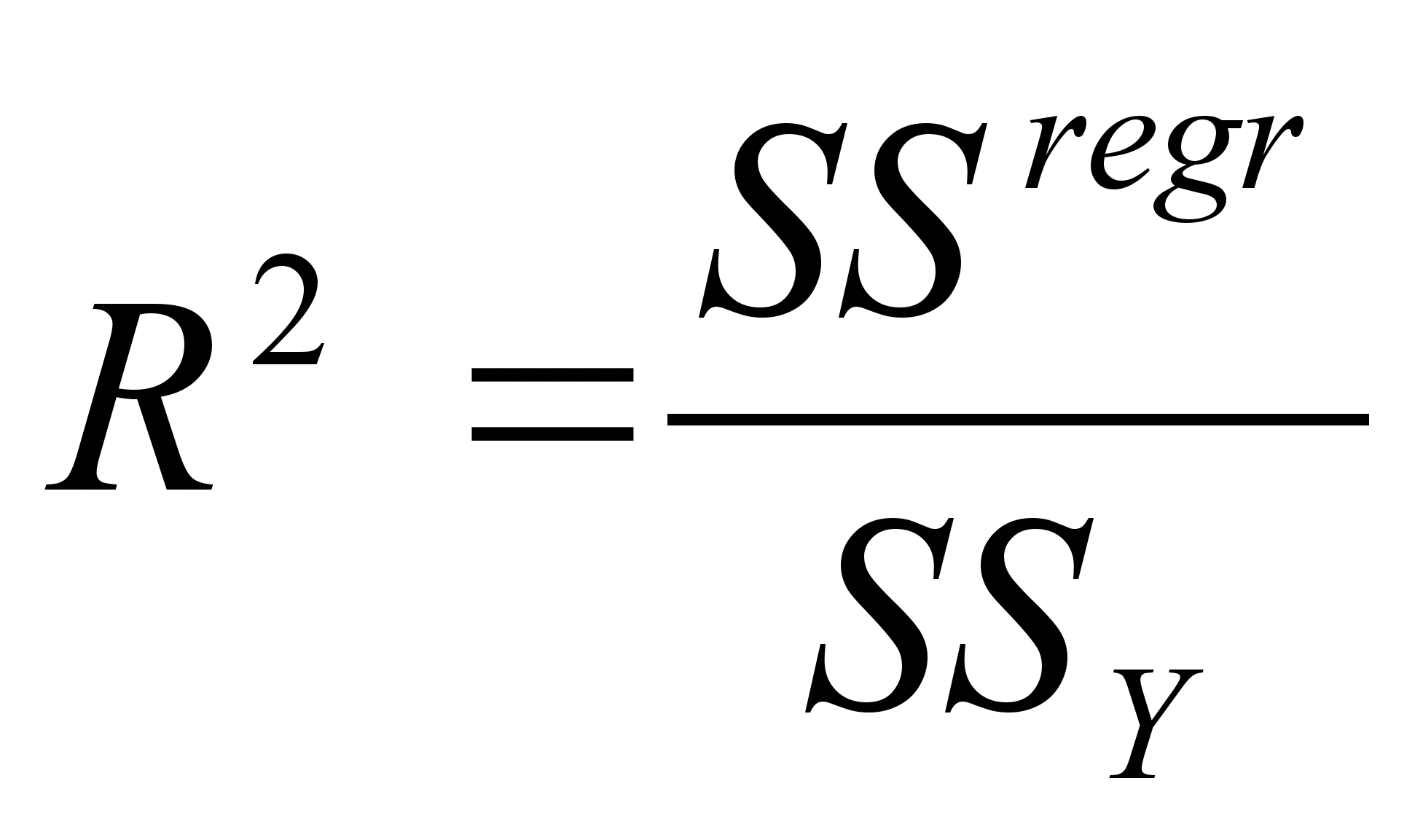
1. Для графического отображения линии регрессии постройте сначала корреляционное поле способом, описанным в задании №1.
2. Далее в окне просмотра результатов  щёлкните дважды на графике, чтобы перенести его в редактор диаграмм.
3. В окне редактора диаграмм выполните команду **Elements/Fit line at total** *(Элементы/Линия тренда для всего графика*).
4. В окне **Properties**(*Свойства*) на вкладке **Fit Line** подтвердите предварительную установку **Linear Regression** и щелкните по кнопке **[Apply]** (*Применить*).
5. Сохраните отчет по заданиям в своей папке под именем *Регресc\_Анализ.spo.*

***Задание 4. Выбор наиболее точной модели связи.***  
  
**Условие задачи**. Исследуется зависимость дозы облучения от толщины слоя защитного материала. Имеются результаты 10 экспериментов (см. рис.24).  
  
Имеются основания предполагать, что зависимость дозы (функция) от толщины  
слоя материала (аргумент) может выражаться одним из следующих уравнений:

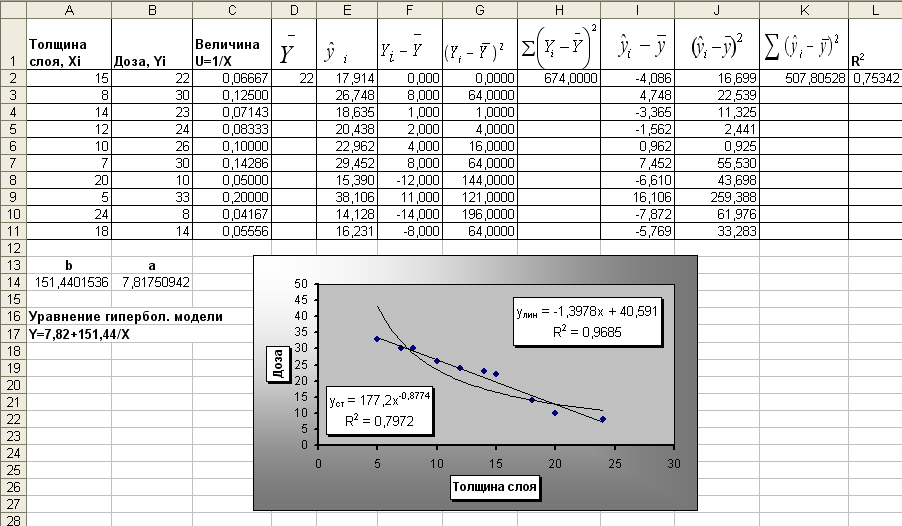
* Y=a b\*X (линейная модель);
* Y=a\*(степенная модель);
* Y=a b/X (гиперболическая модель).

Выберите наиболее точную модель и определите ее коэффициенты.  
  
  
  
  
**Рис.24 Исходные данные для задания №4.**

1. Откройте новую книгу MS Excel и создайте таблицу согласно рис. 25, сохраните в своей папке под именем *Выбор модели.xls*
2. Постройте на этом же листе точечную диаграмму зависимости **Y=f(X).**
3. Нанесите на нее линейный и степенной тренды с уравнениями и величиной  
   достоверности аппроксимации **().**
4. Для построения гиперболической модели преобразуйте модель в линейную, получив в ячейках **С2:С11** величину **1/Х.**
5. Используя функцию ЛИНЕЙН, получите в ячейках **А14:В14** коэффициенты уравнения *b*и *a*(т.е. уравнение).
6. В ячейке **А17**введите уравнение гиперболической модели Y= a b\*U. (вместо *b* и *a*подставьте полученные коэффициенты).
7. Для построенной гиперболической модели найдите величину достоверности  
   аппроксимации **R2.**Для этого сделайте следующие предварительные расчеты:

* Сначала найдите среднее значение  в ячейке **D2.**
* В диапазоне ячеек **E2:E11** получите модельные значения  путем подстановки значений U из блока ячеек **С2:С11** в полученное уравнение гиперболической модели (в ячейке E2 будет формула =$B$14 $A$14\*C2).
* В диапазоне ячеек **F2:F11** рассчитайте (отклонение значений переменной от среднего значения).
* В диапазоне ячеек **G2:G11** рассчитайте  (квадраты отклонений значений переменной от среднего значения).
* В ячейке **H2** найдите сумму квадратов , скорректированную на среднее: .
* Аналогичным образом найдите сумму квадратов прогнозируемых (модельных) значений, обусловленную регрессией. Для этого используйте столбцы **I, J, K.**
* Найдите величину достоверности аппроксимации:  в ячейке **L2.**

1. По значениям коэффициентов достоверности аппроксимации выберите наиболее точную модель, которая соответствует максимальному коэффициенту достоверности.
2. Копия экрана Задания 4 приведена на рис. 25.

  
  
**Рис. 25. Расчеты по выбору оптимальной модели.**