Министерство образования и науки Российской Федерации

Федеральное государственное бюджетное образовательное учреждение высшего

образования «Алтайский государственный технический университет

им. И.И. Ползунова»

**Е. Г. Никифорова**

МЕТОДИЧЕСКИЕ УКАЗАНИЯ К ЛАБОРАТОРНОЙ РАБОТЕ

**«ПАРНАЯ РЕГРЕССИЯ»**

(использование пакета анализа данных)

Изд-во АлтГТУ

Барнаул 2020

ББК 60.822 я 73-1

Никифорова, Е. Г. Методическая разработка «Парная регрессия: использование пакета анализа данных» / Е. Г. Никифорова. – Барнаул: Изд -во АлтГТУ, 2019. – 54 с.

Методическая разработка предназначена для студентов бакалавриата всех форм обучения, изучающих дисциплину «Эконометрика», и написана в соответствии с требованиями соответствующих Федеральных государственных образовательных стандартов.

В методической разработке подробно описан способ построения парных регрессионных моделей с использованием инструментов Excel. Для более эффективного освоения студентами способов применения надстройки «Анализ данных» в разработке содержатся пошаговые иллюстрации действий, а также вопросы для проверки освоения материала.

Протокол № НМС АлтГТУ от 19 апреля 2020 года.

Рецензенты:

Л. П. Афонькина, к.п.н., доцент каф. «Высшая математика» АлтГТУ

Т. Г. Шарикова, к.т.н., доцент каф. «Высшая математика» АлтГТУ

© Е. Г. Никифорова, 2019

© Алтайский государственный технический университет

им. И.И. Ползунова, 2019

**Построение парных регрессионных моделей с использованием пакета «Анализ данных»**

Постановка задачи

Некоторая фирма занимается поставками различных грузов на короткие расстояния внутри города. Требуется оценить время доставки. Помимо расстояния на время доставки, конечно, влияют пробки на дорогах, время суток, дорожные работы, погода, квалификация водителя, вид транспорта, однако в качестве наиболее важного фактора, влияющего на время поставки, менеджер выбрал пройденное расстояние, поскольку в рассматриваемом городе не бывает пробок, в настоящее время не производятся дорожные работы, т.е. все остальные факторы практически не влияют на время доставки.

В таблице приведены данные десяти наблюдений.

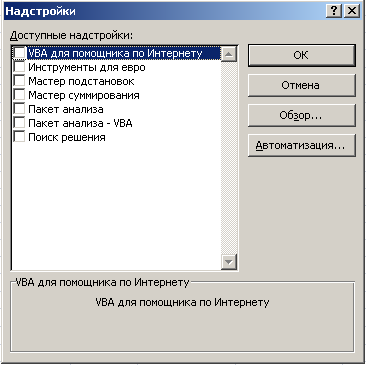
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Время  , мин | 16 | 13 | 19 | 18 | 12 | 11 | 8 | 14 | 9 | 16 |
| Расстояние , км | 3,5 | 2,4 | 4,9 | 4,2 | 3,0 | 1,3 | 1,0 | 3,0 | 1,5 | 4,1 |

Требуется:

1. Построить различные виды парных моделей регрессионной зависимости времени доставки от расстояния: линейную, гиперболическую, степенную, показательную.
2. Вычислить основные качественные характеристики моделей и выбрать лучшую модель, обосновав выбор;
3. Изобразить все линии регрессии на одном графике с полем корреляции.

Решать задачу будем с помощью надстройки над Excel «Анализ данных».Для ее установки откроем файл Excel и найдем надстройки.

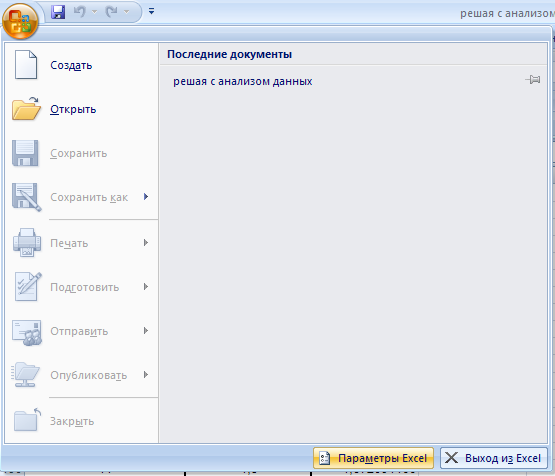
**В WINDOWS-**3 надстройки находятся вменю «Сервис». Нужно нажать «Надстройки», и установить отметку в строке «Пакет анализа»



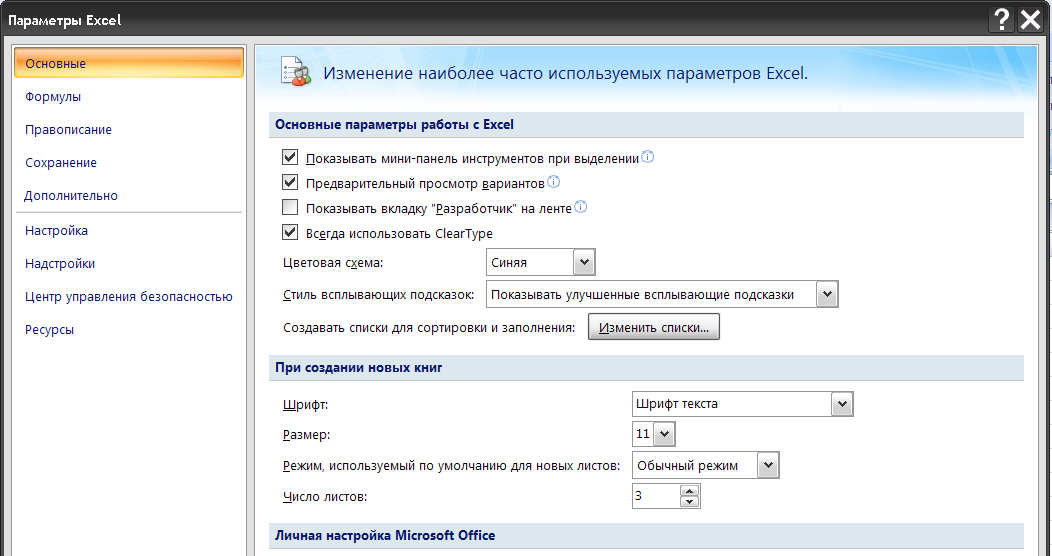
В **WINDOWS-**7 для отыскания настроек нажмем кнопку «Офис» в левом верхнем углу экрана.



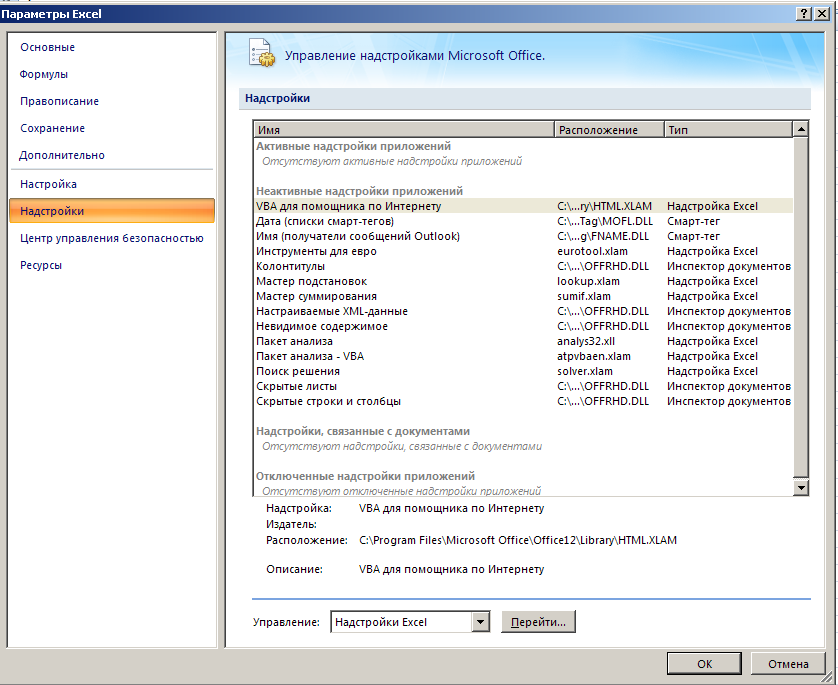
Появится вкладка:



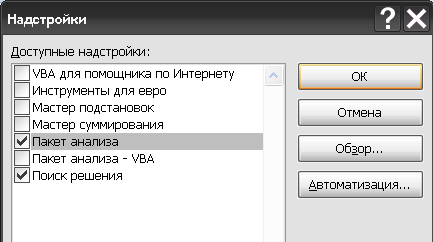
Нажмем «Параметры EXcel», в появившейся вкладке нажмем «Надстройки»



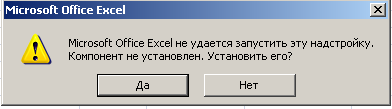
Нажмем «Перейти»



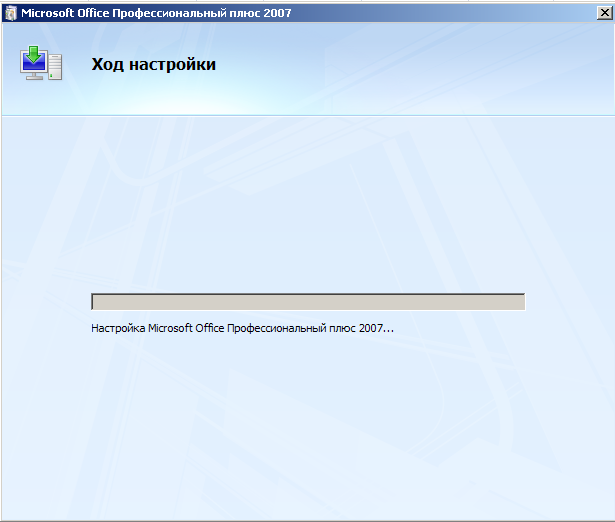
В меню надстроек выберем «Пакет анализа»



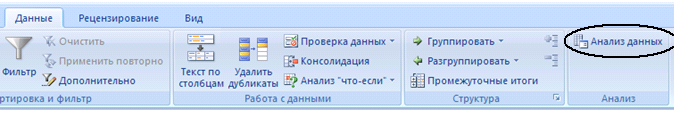
OK



«ДА»



После завершения настройки на вкладке «Данные» появляется раздел «Анализ данных».



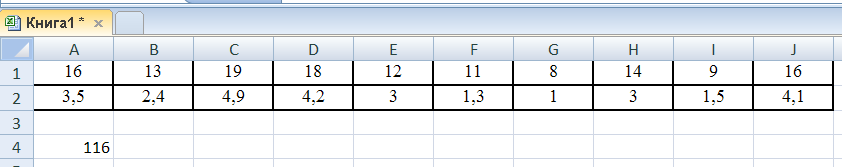
Для любой другой версии «Офиса» нужно найти надстройки, часто они находятся в меню «Файл».

Каждый студент должен обрабатывать свои данные. Они получаются прибавлением числа, которое укажет преподаватель, к данным из таблицы 1. Для примера прибавим число 100.

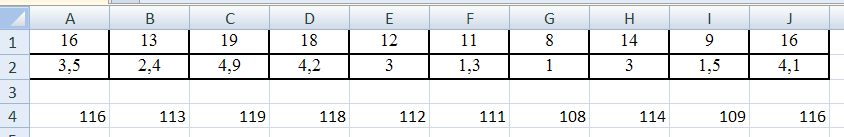
Копируем исходные данные из таблицы.



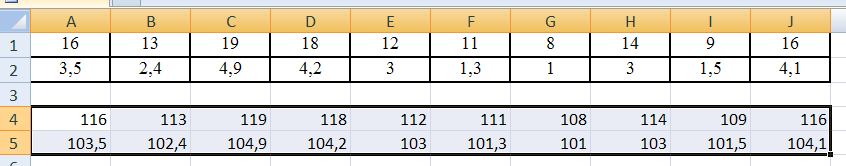
Прибавим число 100. Для этого в ячейке А4 наберем формулу =A1+100. Получим 116:



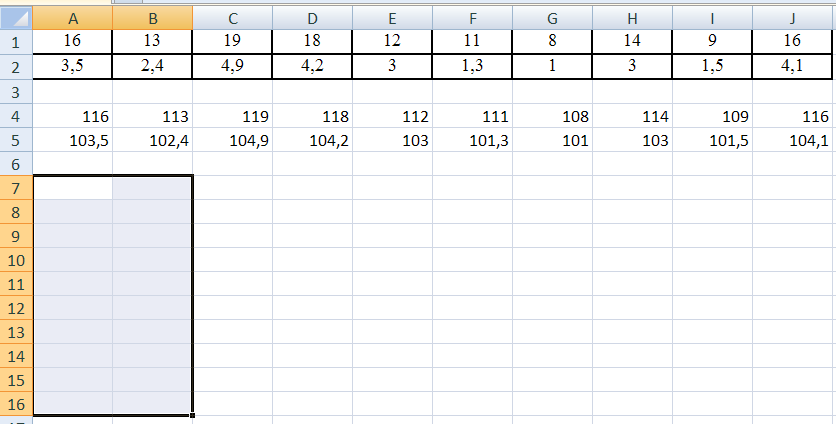
Протянем курсор-крестик в правом нижнем углу ячейки А4 вправо



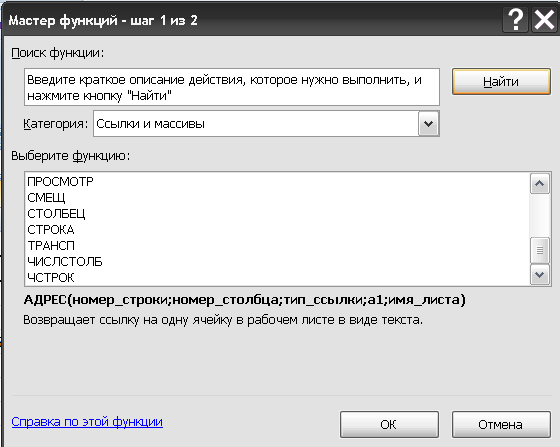
и затем вниз:



Для того, чтобы данные располагались в столбцах, полученный массив нужно транспонировать. Для этого сначала на листе Excel выделим место, где будет располагаться транспонированный массив.

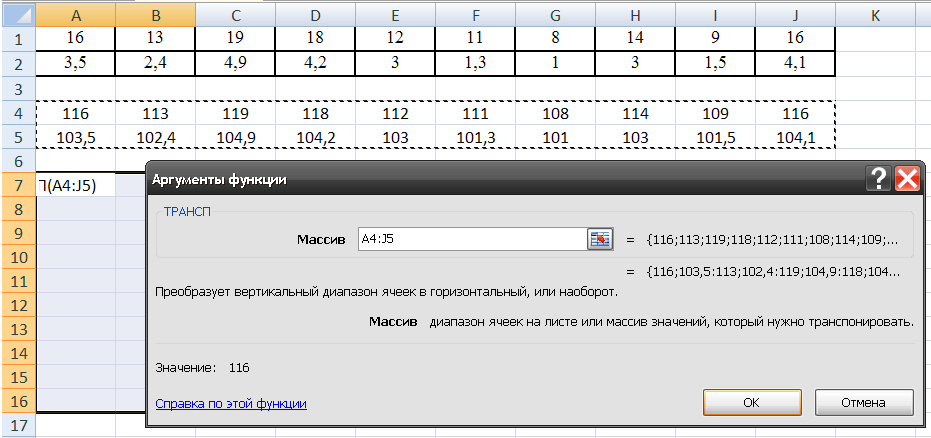


Нажмем кнопку «Вставка функции» , выберем категорию «Ссылки и массивы», а в ней функцию ТРАНСП



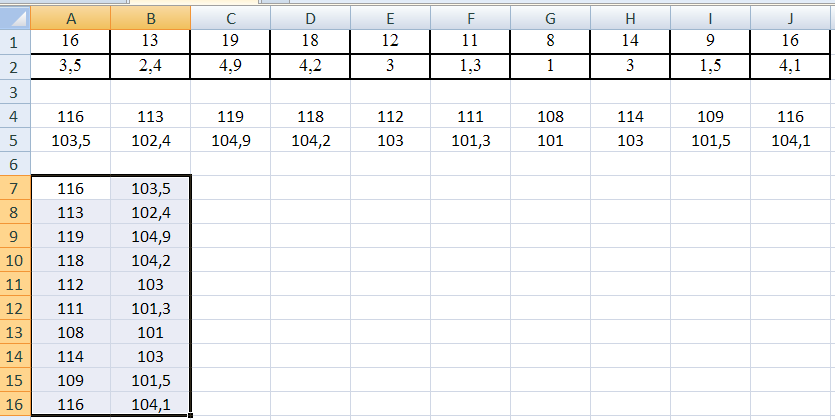
ОК

В открывшемся диалоге отметим массив, который требуется транспонировать

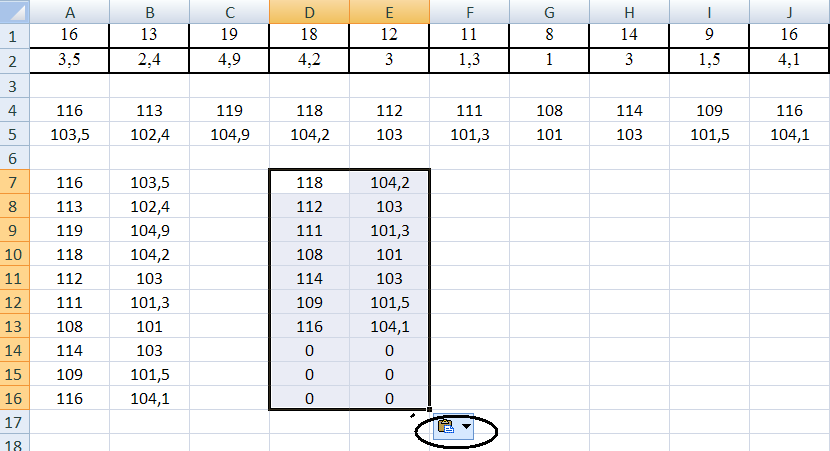


и нажмем одновременно «Ctrl-Shift-Enter»

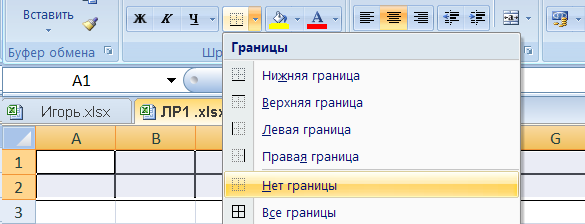
Результат:



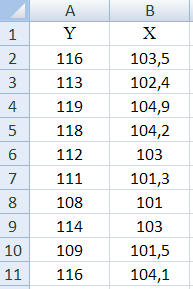
Теперь нужно убрать с листа все лишнее. Копируем полученные колонки и вставляем рядом



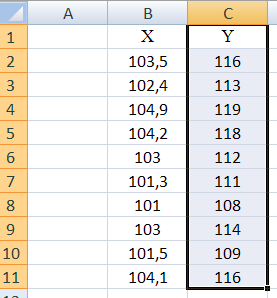
Чтобы значения копировались правильно, зайдем в «Параметры вставки» (на рисунке отмечено овалом). В открывшемся окне выберем «Только значения». После этого на листе можно убрать все лишнее. Рамки таблицы уберем, выделив на листе таблицу и нажав в меню «Границы» строку «Нет границы»:



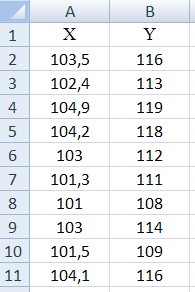
Сдвинем данные в левый верхний угол листа и озаглавим столбцы



Переставим столбцы, чтобы было удобно работать в дальнейшем. Для этого выделим первый столбец, поставим курсор на границу, получим резной крестик и перенесем столбец в нужное место:

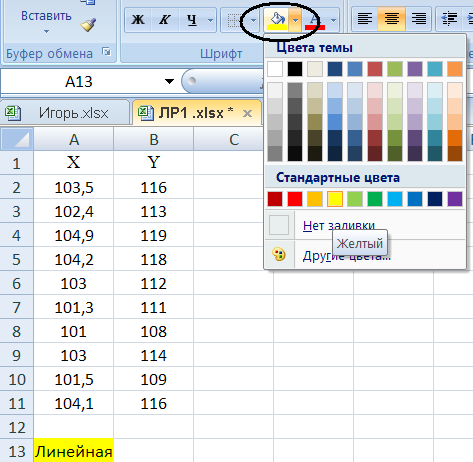


Далее, выделим оба столбца, и аналогично перенесем весь массив в левый верхний угол листа.

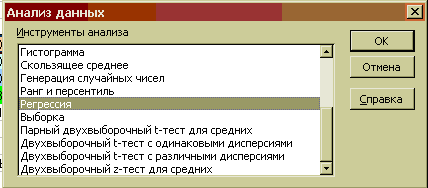


**Построение линейной модели**

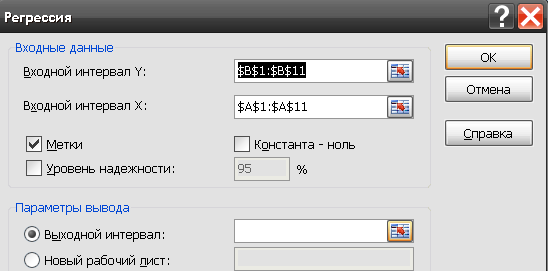
Покрасим ячейку А13 в желтый цвет и подпишем название первой модели – «Линейная»:



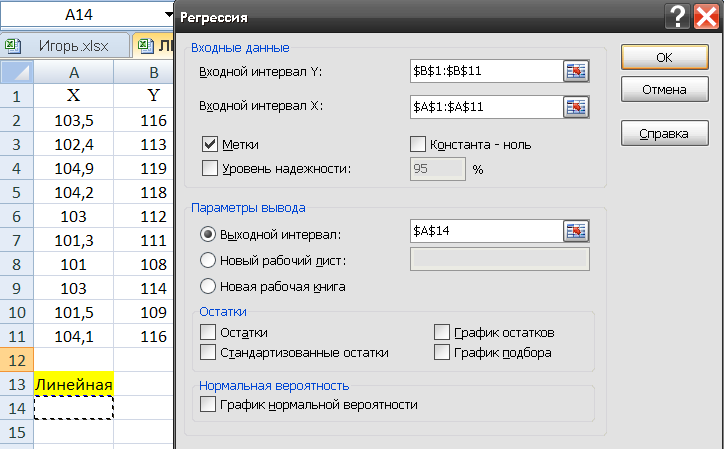
На вкладке «Данные» выберем «Анализ данных», затем «Регрессия» и нажмем OK



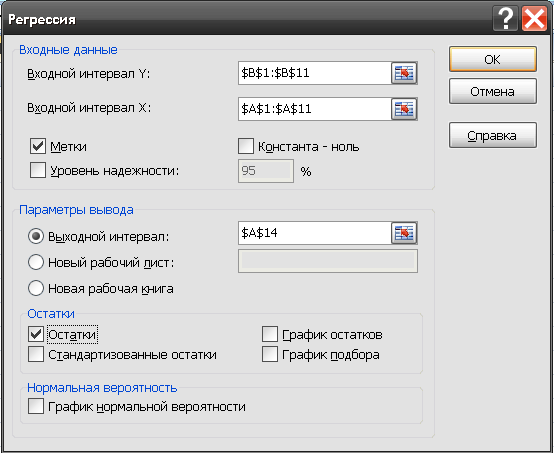
В появившемся диалоговом окне щелкнем курсором в строке «Входной интервал Y» и на листе EXcel выделим колонку Y вместе с заголовком. Затем в строке «Входной интервал X» выделим колонку X вместе с заголовком, в строке «Метки» поставим «Галочку», что будет сигнализировать машине о том, что в первой строке находятся названия данных, а не сами данные, щелкнем курсором на кнопку «Выходной интервал». При этом выделение перейдет в строку «Входной интервал Y»:



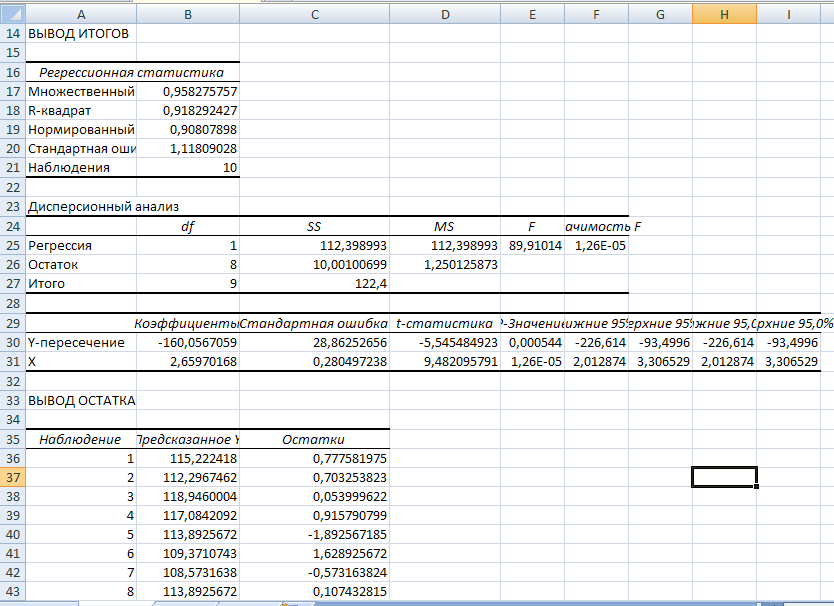
Нужно вернуть его в строку «Выходной интервал», щелкнув курсором по его полю. В качестве выходного интервала укажем ячейку А14, щелкнув по ней курсором:



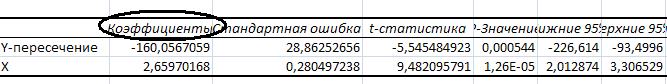
Поставим отметку в строке «Остатки» и нажмем ОК.



Получим вывод итогов:



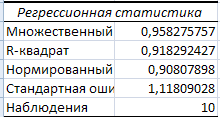
Коэффициенты модели находятся в столбце «Коэффициенты»:



в строке «Y-пересечение» записан коэффициент , он равен -160,056059, ниже – коэффициент . Согласно принятому в статистике соглашению, к точности исходных данных в промежуточных вычислениях добавляют две значащих цифры, а в результате одну значащую цифру, таким образом, уравнение регрессии имеет вид: .

Кроме коэффициентов уравнения регрессии в таблице «Вывод итогов» содержится следующая информация (далее все значения приводятся с учетом округления).

Раздел «Регрессионная статистика»:



коэффициент детерминации  (вторая строка);

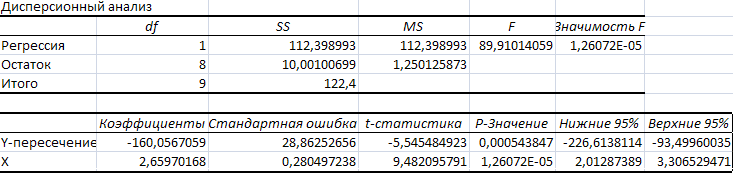
индекс детерминации  (первая строка);

нормированный коэффициент детерминации  (третья строка);

стандартная ошибка 

количество наблюдений .

Раздел «Дисперсионный анализ»:



 – число степеней свободы, при этом:

 – число степеней свободы общей дисперсии;

 – число степеней свободы факторной дисперсии;

 – число степеней свободы остаточной дисперсии.

 – сумма квадратов:

- общая сумма квадратов; 

 - факторная сумма квадратов;

- остаточная сумма квадратов.

 – сумма квадратов, рассчитанная на одну степень свободы (т.е. сумма квадратов, деленная на степень свободы:

 - для факторной суммы квадратов;

- для остаточной суммы квадратов.

Значение F-критерия: .

Значения -статистик ; ;

стандартные ошибки вычисления t-статистик , ;

95%-ные доверительные интервалы для параметров  и  уравнения регрессии (интервалы, накрывающие неизвестный параметр с вероятностью 0,95) , ;

*P-значение* – характеристика значимости соответствующего коэффициента уравнения регрессии: если *P*-значение мало, то найденный коэффициент значимо отличается от нуля и его можно использовать. В рассматриваемой модели , что говорит о значимости каждого коэффициента уравнения линейной регрессии.

Итак, уравнение регрессии имеет вид: 

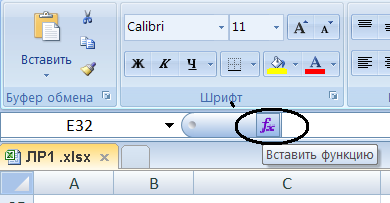
Экономический смысл коэффициента регрессии : с увеличением расстояния доставки на 1км среднее время доставки увеличивается на 2,66 минуты.

Коэффициент  в данной модели экономического смысла не имеет.

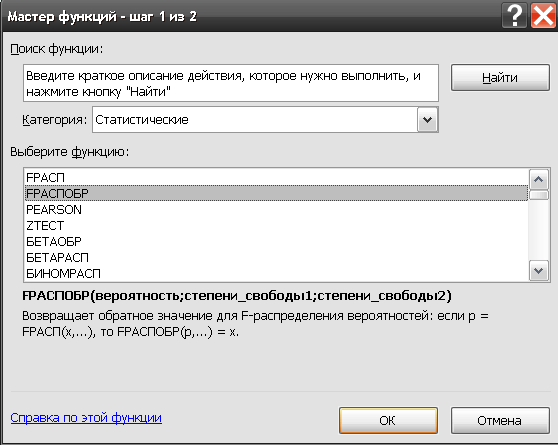
Коэффициент детерминации  говорит о том, что модель адекватна (поскольку ), изменение среднего времени доставки на 91,8% объясняется изменением дальности доставки.

Индекс детерминации , что говорит о сильной зависимости между средним временем доставки и расстоянием доставки.

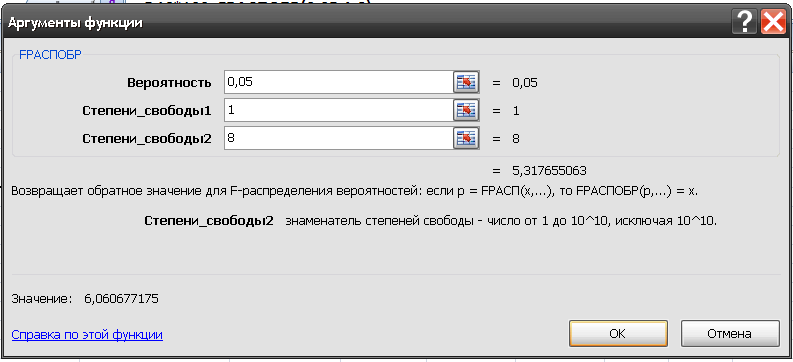
Наблюдаемое значение F-критерия . Найдем критическое значение F-критерия. Для этого поставим курсор в ячейку E23, затем нажмем кнопку  «Вставка функции».



В категории «Статистические» выберем функцию FРАСПОБР и нажмем ОК.



В строке «Вероятность» укажем 0,05 (это уровень значимости); в строке «Степени свободы 1» укажем количество независимых переменных в модели – 1; «Степени свободы 2» вычислим по формуле ; нажмем ОК.

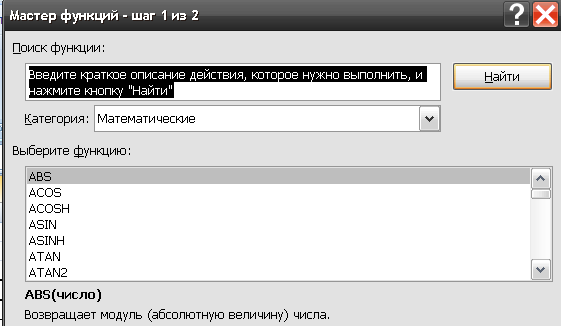


В результате получим критическое значение F-критерия: . Поскольку, то уравнение регрессии в целом и показатель тесноты связи статистически значимы и надежны.

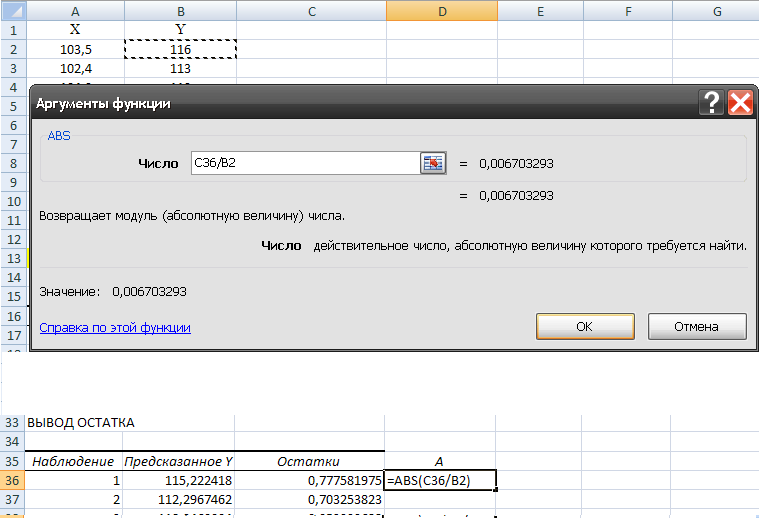
Вычислим среднюю ошибку аппроксимации . Величины  составляют разницу (ошибку предсказания) между исходным значением y и значением y, вычисленным по уравнению регрессии, т.е. предсказанным значением y) и содержатся в выводе итогов в разделе «Вывод остатка», колонка называется «Остатки».



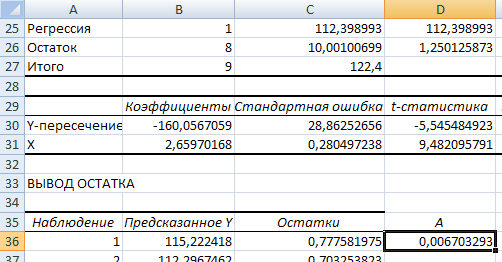
В ячейке D36 вызовем вставку функции, в категории «Математические» выберем функцию ABS



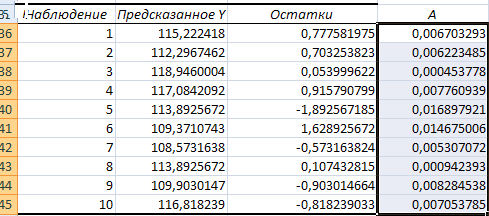
и укажем, что модуль берем от дроби C36/B2. Нажмем ОК.



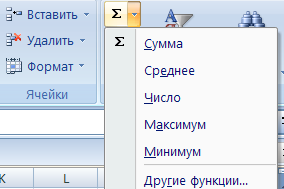
В ячейке D36 появится значение 0,006703293



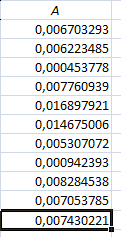
Протянем курсор-крестик вниз и получим значения модулей относительных ошибок аппроксимации.



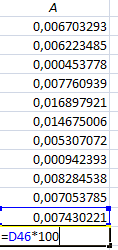
Найдем среднее значение в колонке А (для вывода меню нажмем значок ▼ на кнопке.

****

Результат:



Для получения средней ошибки аппроксимации нужно полученный результат умножить на 100:

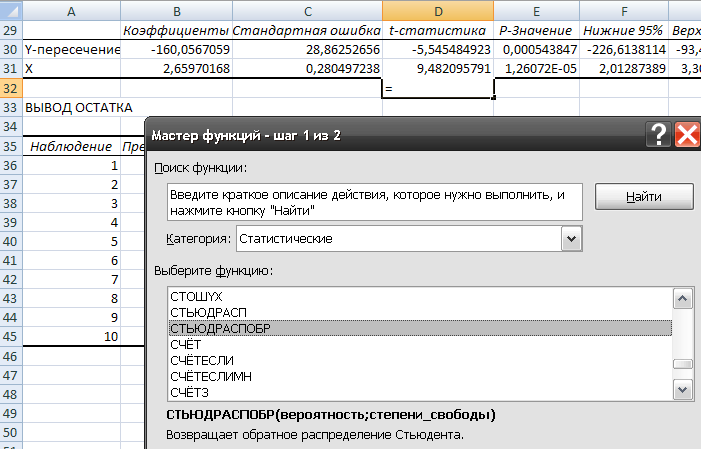


Итак,

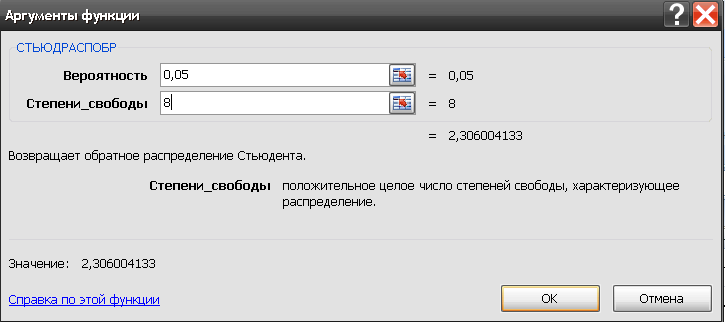
.

Таким образом, средняя ошибка аппроксимации построенной модели составляет 0,743%, что меньше 10%, поэтому модель точная, причем вычисленные значения y отличаются от исходных данных в среднем на 0,743%.

Для проверки статистической значимости и надежности каждого коэффициента уравнения регрессии сравним значения t-статистик  и  с критическим значением . В ячейке D32 щелкнем вставку функции – кнопку , категория «Статистические», функция СТЬЮДРАСПОБР. НажмемOK.



В диалоге «Аргументы функции» укажем вероятность 0,05, степени свободы вычислим по формуле 

****

Итак, 2,306

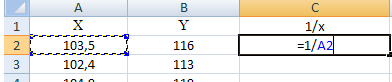
Поскольку , , то коэффициенты  и уравнения регрессии статистически значимы и надежны.

Укажем - 95% доверительные интервалы для параметров  и  уравнения регрессии. Нижняя и верхняя границы интервалов содержатся в выводе итогов в таблице «Дисперсионный анализ», колонки называются *Нижние 95%* и *Верхние 95%.* Итак, , с вероятностью 0,95, с учетом округления: , .

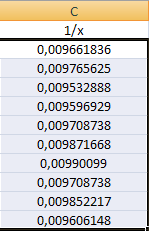
**Построение гиперболической модели**

Для построения гиперболической модели вида  рассмотренным способом озаглавим новую колонку  и вычислим в ней соответствующие значения. В ячейке А2 наберем формулу

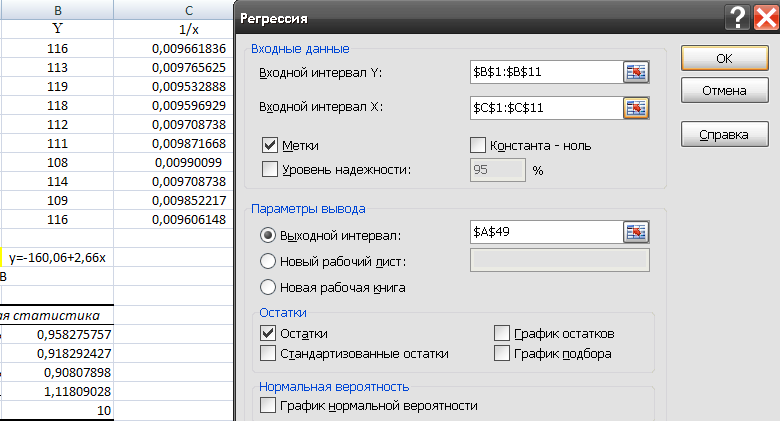
=1/A2



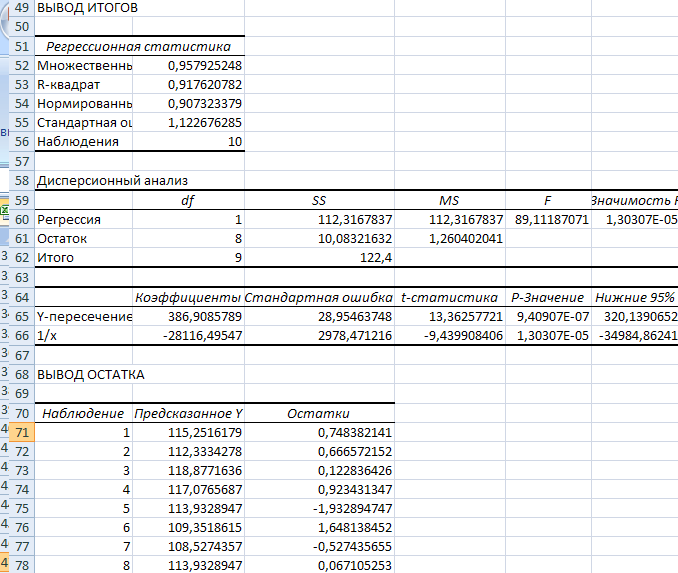
Получив результат 0,009661836, протянем курсор вниз и получим все значения



Покрасим ячейку А48 в желтый цвет и подпишем «Гиперболическая». Воспользуемся пакетом «Анализ данных», инструмент «Регрессия». В диалоге окна «Регрессия» в строке «Входной интервал Y»ничего не изменилось по сравнению с линейной моделью, там выделена колонка Y вместе с названием. В строке «Входной интервал X» выделяем вместе с названием колонку . Изменится выходной интервал, теперь это ячейка А49.



Результат:



Таким образом, уравнение регрессии имеет вид: .

Коэффициент детерминации , модель адекватна, изменение среднего времени доставки на 91,76% изменением дальности доставки.

Индекс детерминации , что говорит о сильной зависимости между средним временем доставки и расстоянием доставки.

Наблюдаемое значение F-критерия . Поскольку

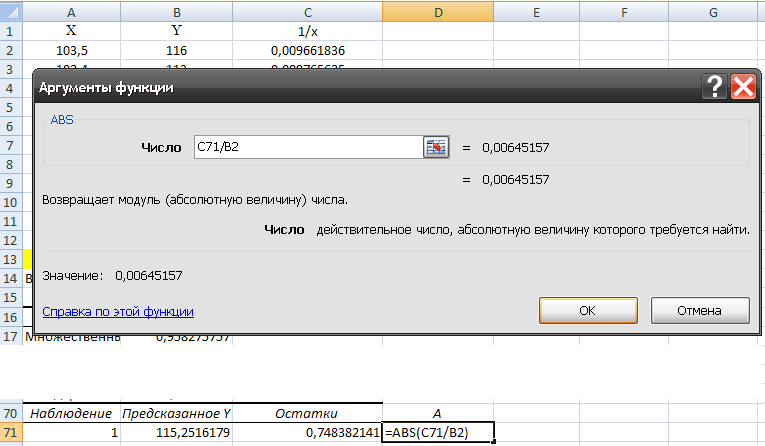
,

то уравнение регрессии в целом и коэффициент детерминации являются статистически значимыми и надежными.

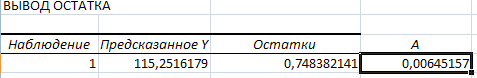
Вычислим среднюю ошибку аппроксимации по формуле

,

используя вывод остатков. Щелкнем курсором в ячейке D71, затем щелкнем кнопку  «Вставка функции»: выберем категорию математические, в ней функцию ABS, щелкнем ячейку С38, затем знак деления / и ячейку В2.

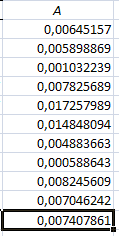


Результат:

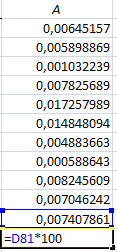


Протянем курсор вниз и, получив все значения в колонке A, вычислим их среднее, нажав значок ▼ на кнопке 

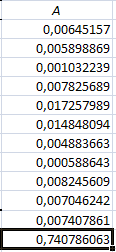
В результате имеем:



Значение  получается умножением полученного результата на 100.



Получим



Таким образом,



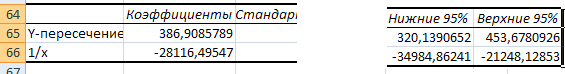
Итак, средняя ошибка аппроксимации построенной модели составляет %, модель точная, вычисленные значения отличаются от исходных данных в среднем на %.

Поскольку

,

то коэффициенты  и  в уравнении являются статистически значимыми и надежными.

Укажем 95% доверительные интервалы для параметров  и  уравнения регрессии. Поскольку



то

, 

Внесем все полученные показатели в итоговую таблицу:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| название | уравнение | R2 | A | F |  |  |
| линейная |  | 0,918 | 12,91 | 89,910 | -3,531 | 9,482 |
| гиперболическая |  | 0,848 | 21,43 | 44,530 | 10,996 | -6,673 |
| степенная |  |  |  |  |  |  |
| показательная |  |  |  |  |  |  |

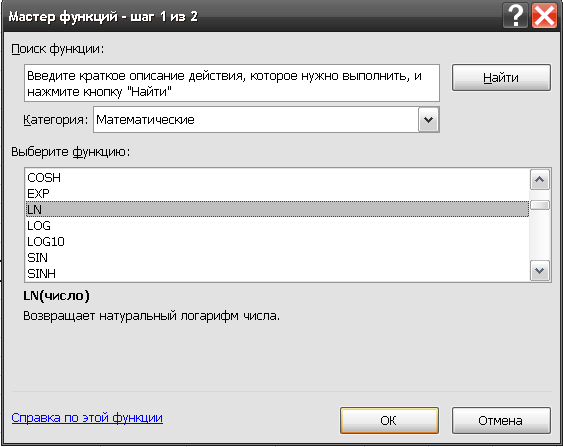
**Построение степенной модели**

Для построения модели вида  рассмотренным способом прологарифмируем исходное уравнение:

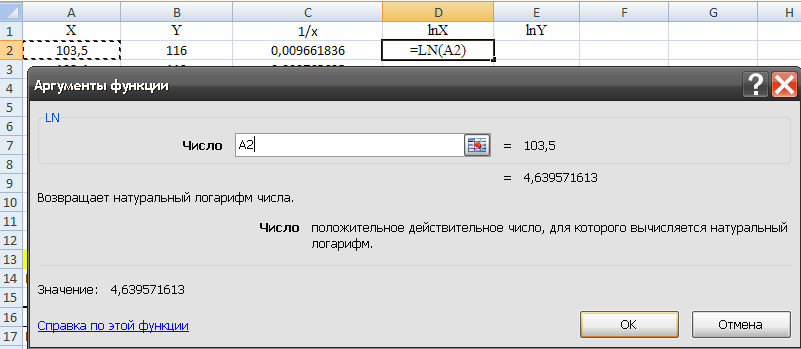
,

Вычислим lny и .

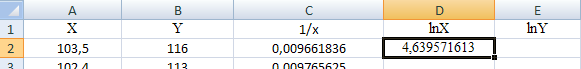
Нажмем – вставку функции, выберем категорию «математические», функция LN нажмем OK.



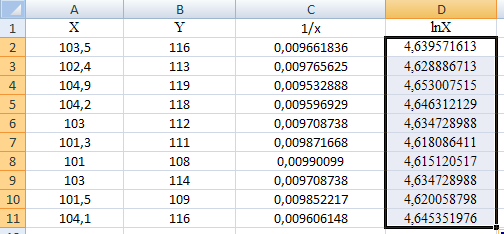
В появившемся диалоге в строке «Число» щелкнем ячейку А2 – первое значение X.



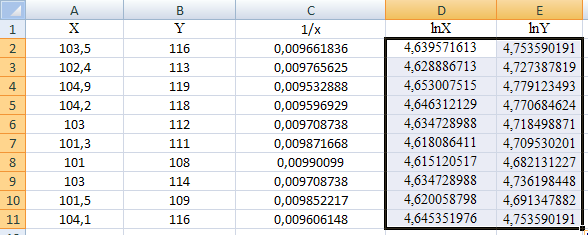
Нажмем ОК, получим результат:



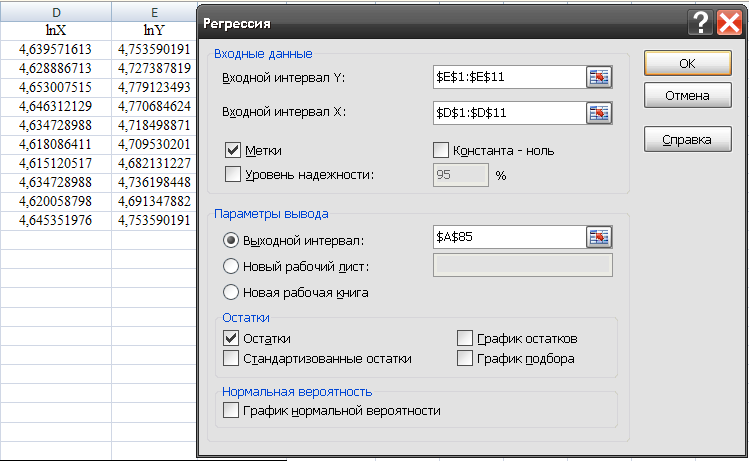
Протянем курсор вниз и получим все значения.



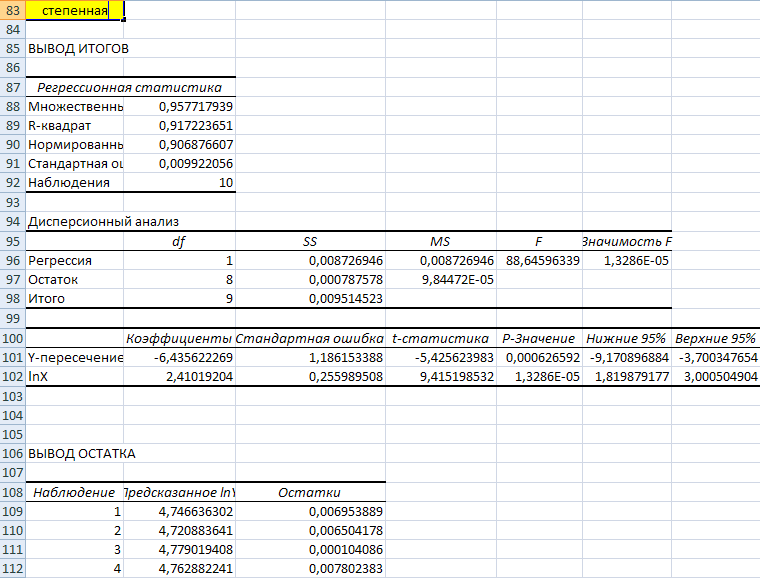
Для получения логарифмов Y, сдвинем курсор вправо в соседнюю колонку.



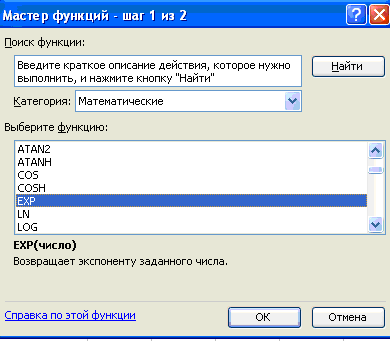
Щелкнем курсором в любую свободную ячейку листа. Далее используем пакет анализа, инструмент «регрессия», где в качестве входного интервала Y укажем колонку lnY, а в качестве входного интервала X – колонку lnX, выходной интервал – ячейка A85.



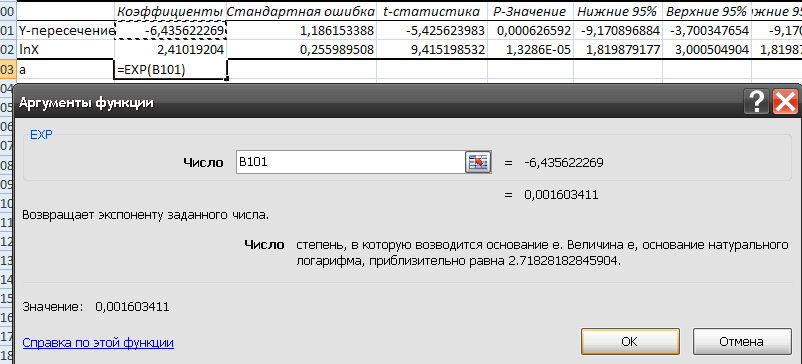
Получим следующий вывод итогов:



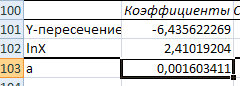
Поскольку для линеаризации степенного уравнения его пришлось логарифмировать, т.е. фактически, мы работаем с уравнением , то в строке «Y-пересечение» находится значение . Для получения значения  нужно выполнить потенцирование, т.е.к значению  применить функцию exp (поскольку ) . В ячейку B103 вызовем «Вставку функции» , выберем категорию «Математические», функция EXP, и нажмем OK.



В качестве аргумента функции укажем (щелкнем) ячейку А101.



Итак, .



Запишем уравнение степенной модели : y=0,0016\*X^2,41, т.е. .

В степенной модели коэффициент *b* имеет экономический смысл – это коэффициент эластичности. При увеличении  на 1% значение  увеличивается, в среднем на 2,41%

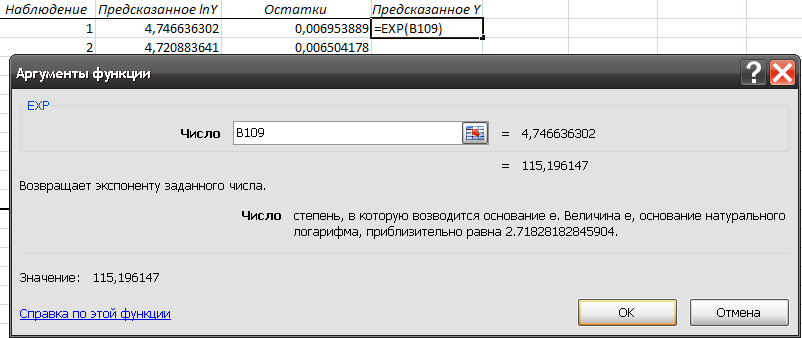
Коэффициент детерминации , модель адекватна, изменение среднего времени доставки на 91,7% объясняется изменением дальности доставки.

Наблюдаемое значение F-критерия  68,3763, уравнение регрессии в целом и коэффициент детерминации статистически значимы и надежны.

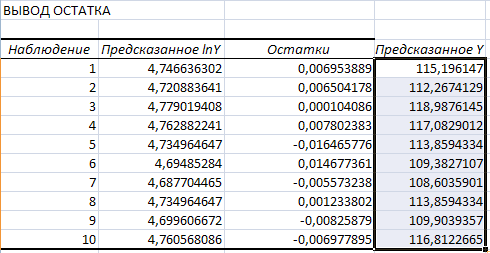
Вычислим среднюю ошибку аппроксимации

.

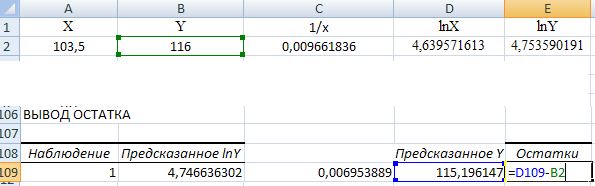
Поскольку в выводе остатка степенной модели представлено «*Предсказанное lnY»,* введем столбец «*Предсказанное Y*», в котором вычислим экспоненты от элементов столбца «*Предсказанное lnY».*



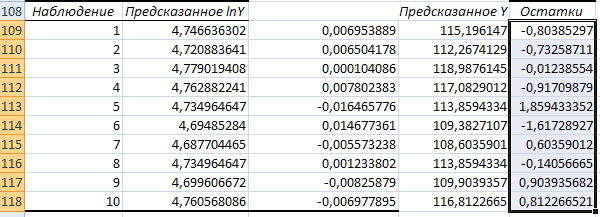
Результат:



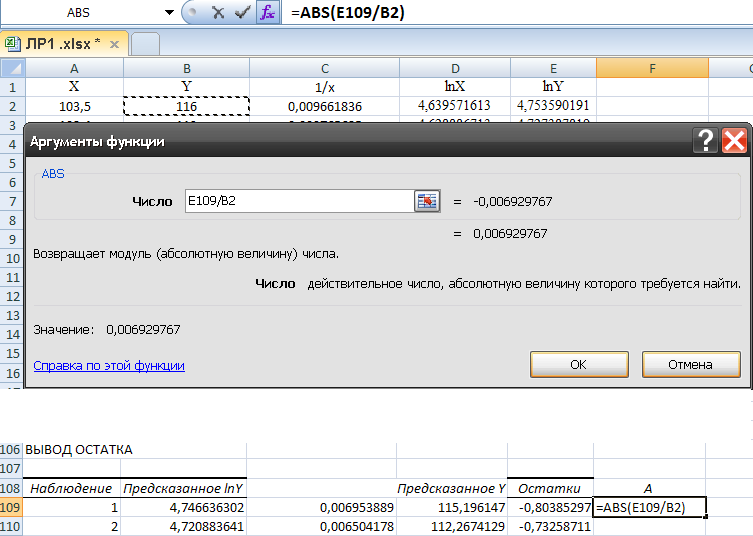
В соседнем столбце вычислим остатки , т.е. от предсказанного значения y отнимем исходное значение.



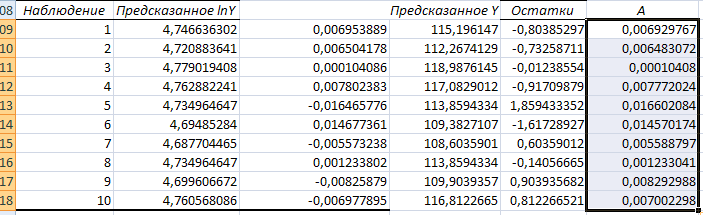
Вычислим все значения столбца:

****

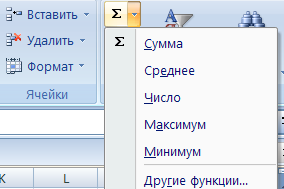
Вычислим среднюю ошибку аппроксимации в соседней колонке.



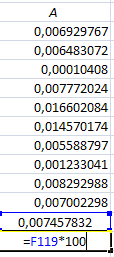
Результат:



Для вычисления среднего по столбцу в меню значка «Сумма» щелкнем «Среднее».

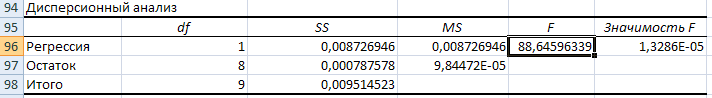


Полученный результат умножим на 100.



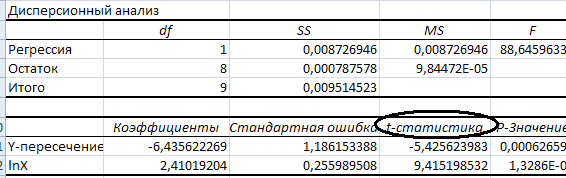
Итак, . Таким образом, ошибка построенной модели меньше одного процента, что говорит о высокой точности модели, вычисленные значения  отличаются от исходных данных в среднем на %.

Значение F-критерия берем из таблицы вывода итогов



Поскольку , то уравнение регрессии в целом и коэффициент детерминации статистически значимы и надежны.

Значения t-статистик берем из таблицы вывода итогов.

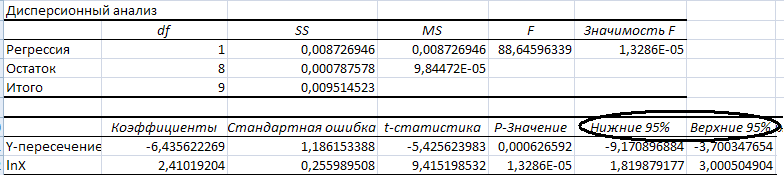


Поскольку

,

то коэффициенты  и  в уравнении статистически значимы и надежны.

Укажем границы 95% доверительных интервалов для параметров уравнения регрессии.





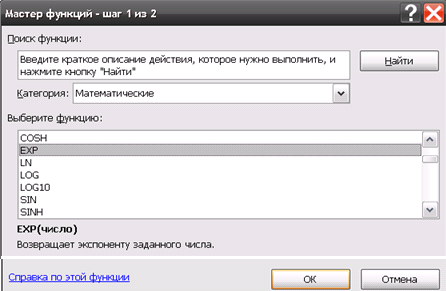
Поскольку в таблице вывода итогов в строке «Y-пересечение» указано значение , то в колонках «*Нижние 95%»* и *«Верхние 95%»* указаны границы изменения :



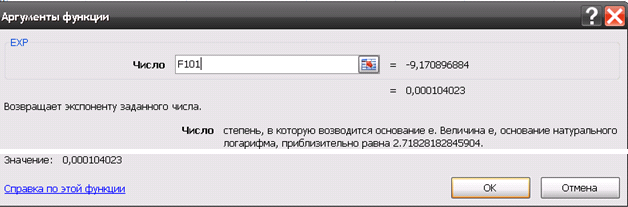
Тогда

 или .

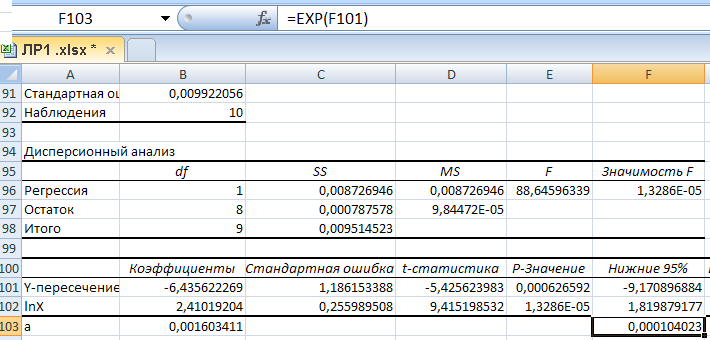
Таким образом, для получения границ изменения коэффициента , нужно применить экспоненту к указанным границам. В ячейке F103 щелкнем значок «Вставка функции» и выберем функцию EXP.



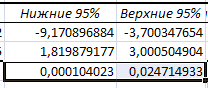
В качестве аргумента экспоненты укажем ячейку F101.



Получим значение 0,000104023 левого конца интервала для параметра  уравнения регрессии



Протянем курсор-крестик вправо и получим значение правого конца интервала.



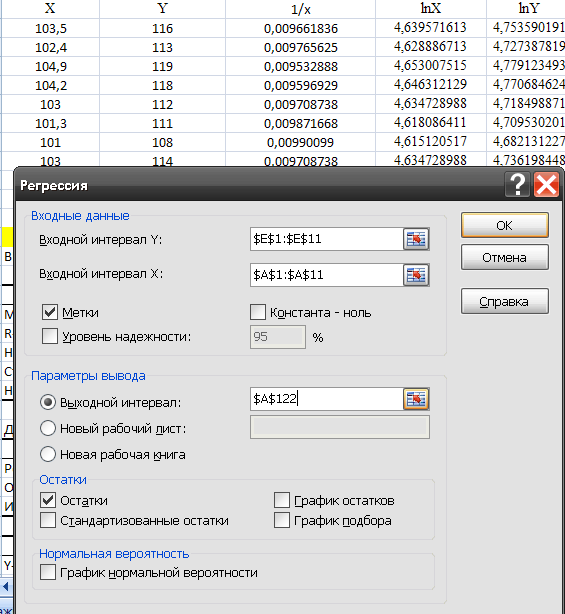
Итак, .

**Построение показательной модели**

Для построения модели вида  линеаризуем уравнение, применив логарифмирование:



Воспользуемся пакетом анализа данных, функцией «Регрессия». В строке «Входной интервал Y» выделим колонку lnY, где в строке «Входной интервал X» выделим колонку X, поставим отметку в строках «Метки» и «Остатки», выходной интервал – ячейкаА122.

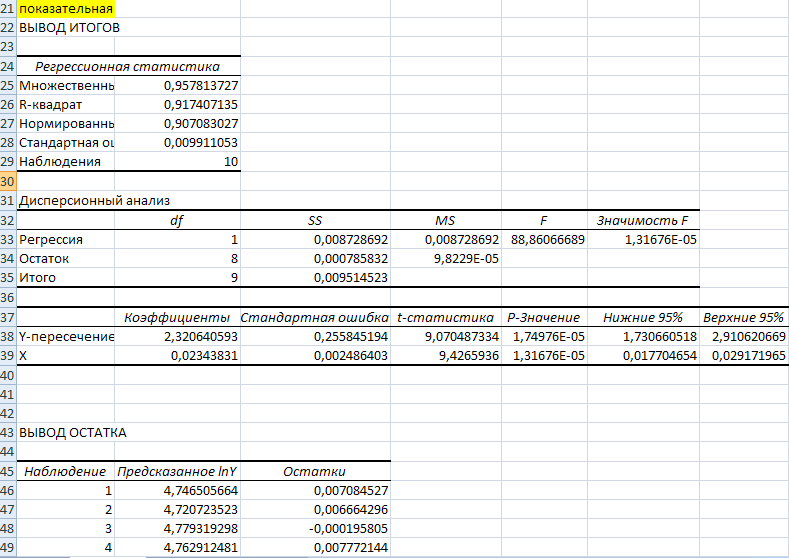


После нажатия OK получим вывод итогов.

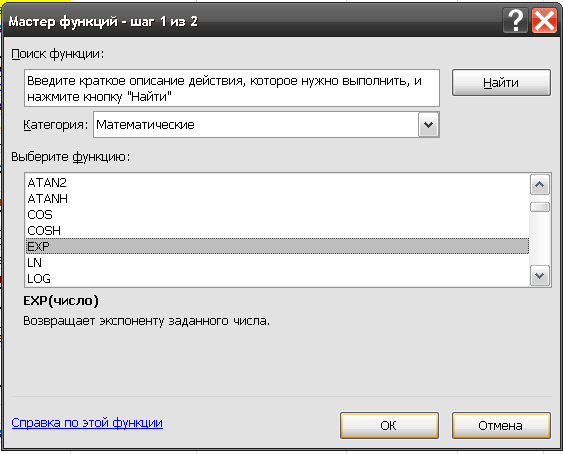
Заметим, что из-за логарифмирования, в строке «Y-пересечение» находится , а в строке «X» указан , а в колонках «*Нижние 95%» и «Верхние 95%»* находятся верхние и нижние границы доверительных интервалов  и  соответственно, т.е.

.

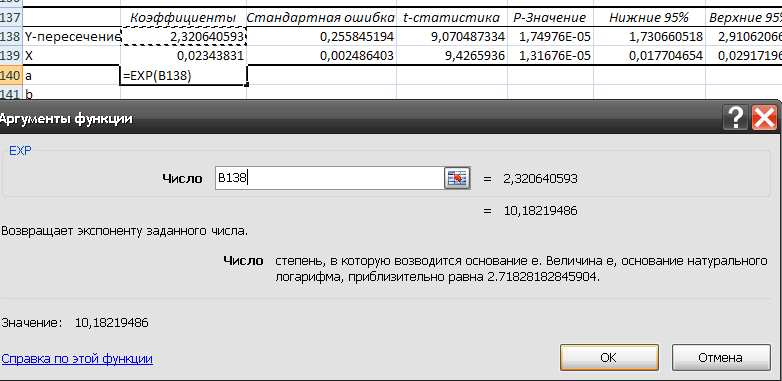
Нам же требуются значения самих  и , а также границы их доверительных интервалов.



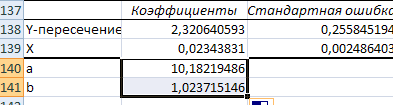
В ячейку В140 вызовем вставку функции  в категории «Математические», выберем функцию EXP, OK.



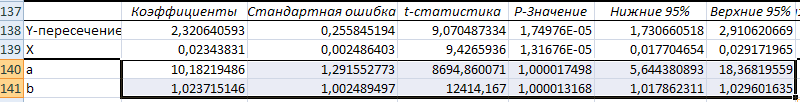
В качестве аргумента укажем ячейку В138.



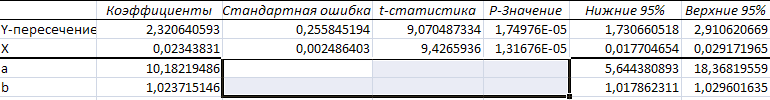
Получим результат 10,18219486 и протянем курсор сначала вниз на одну ячейку,



а затем вправо до колонок «*Нижние 95%*» и «*Верхние 95%*» включительно.



Убрав лишние значения, имеем



Итак,

,

.

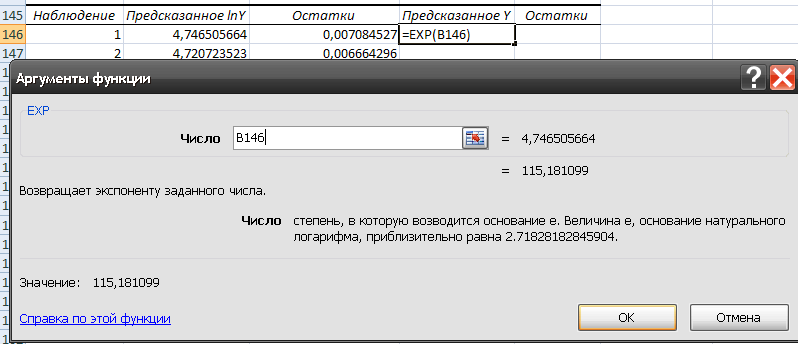
Уравнение модели: .

Коэффициент детерминации , значит, модель адекватна, изменение среднего времени доставки на 91,7% объясняется изменением дальности доставки.

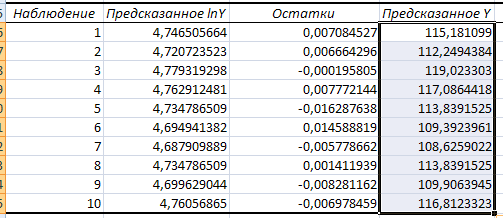
Индекс детерминации , что свидетельствует о сильной связи между средним временем доставки и расстоянием.

Наблюдаемое значение F-критерия: , значит, уравнение регрессии в целом и коэффициент детерминации статистически значимы и надежны.

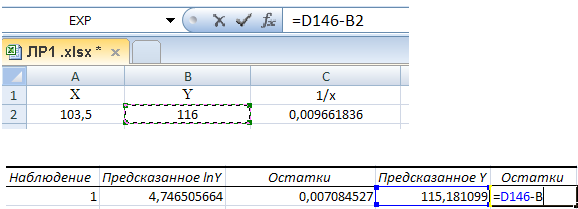
Для вычисления средней ошибкиаппроксимации по формуле  найдем предсказанные значения y и остатки. Для этого в ячейку D146 вызовем функцию EXP и в качестве аргумента укажем первое значение предсказанного lnY.



Получим все значения y:



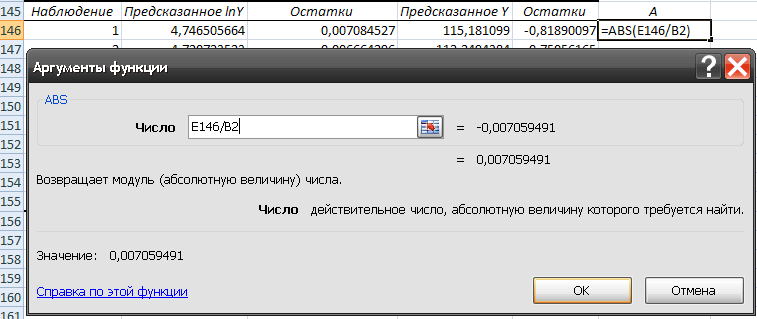
Вычислим остатки.



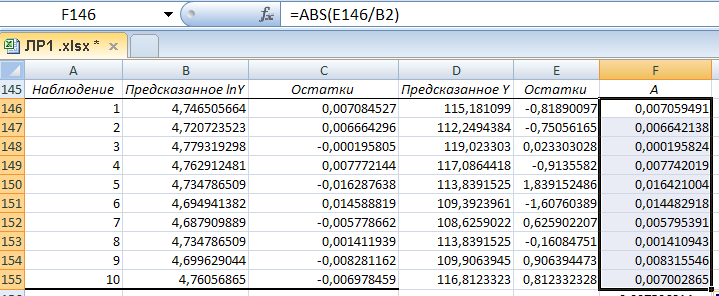
Все значения остатков:



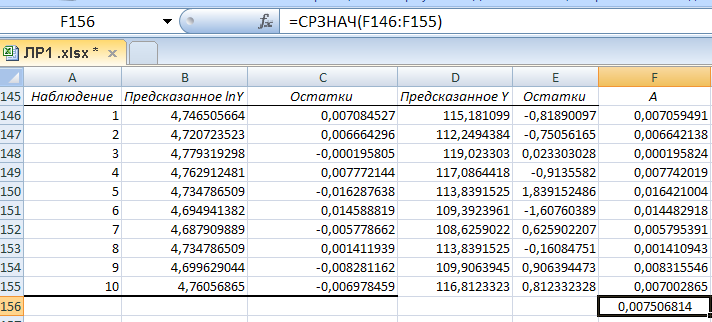
Вычислим среднюю ошибку аппроксимации .



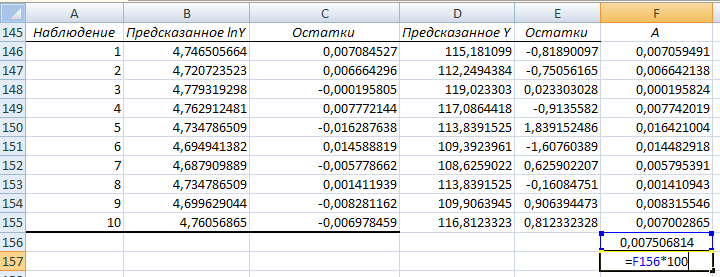
Результат:



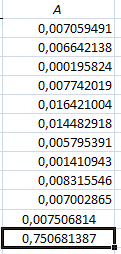
Среднее значение по колонке:



Вычислим значение :

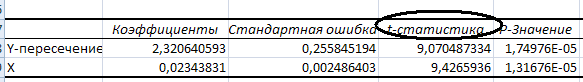


Результат:



Таким образом, , модель точная, вычисленные значения отличаются от исходных данных в среднем на 0,75%.

Значения t-статистик берем из «Вывода итогов».

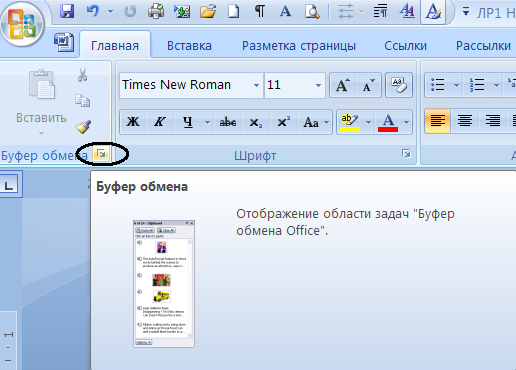


Поскольку

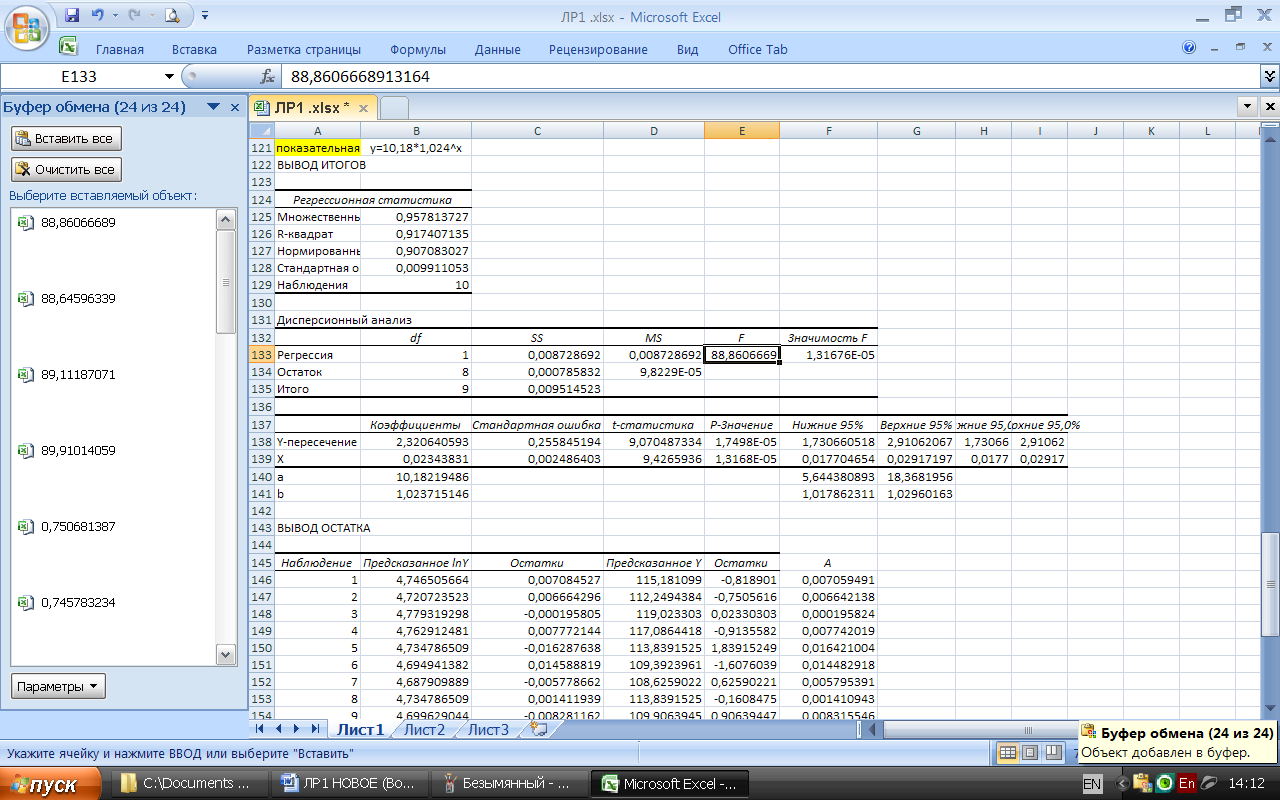
,

То коэффициенты модели статистически значимы и надежны.

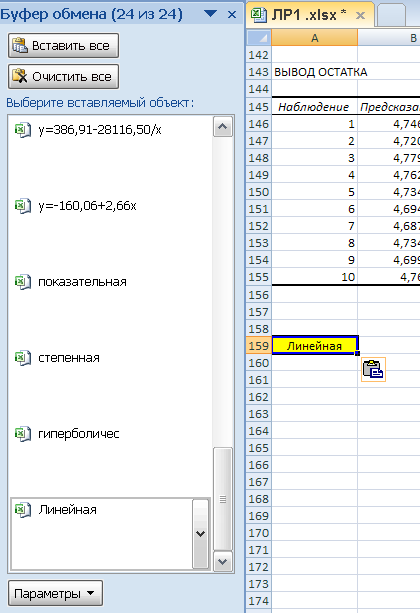
Внесем все полученные показатели в итоговую таблицу. Для этого на вкладке «Главная» откроем «Буфер обмена»,



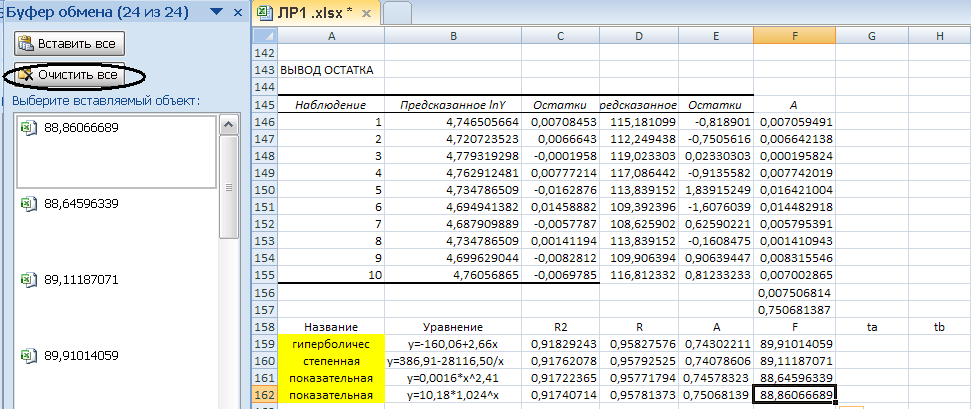
Затем поочередно копируем названия уравнений, формулы зависимостей, коэффициенты детерминации, индексы детерминации, значения средних ошибок аппроксимации и F-критериев. Буфер обмена заполнен полностью: 24 записи из 24 возможных.



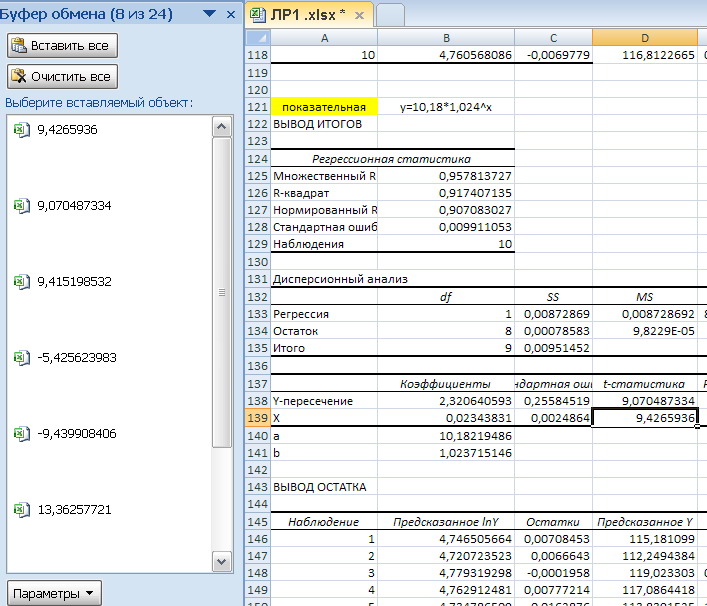
На свободном месте листа EXcel вставим записи в таблицу, начиная с первой. (Для этого бегунок буфера обмена опустим вниз и щелкнем первую запись «Линейная».)



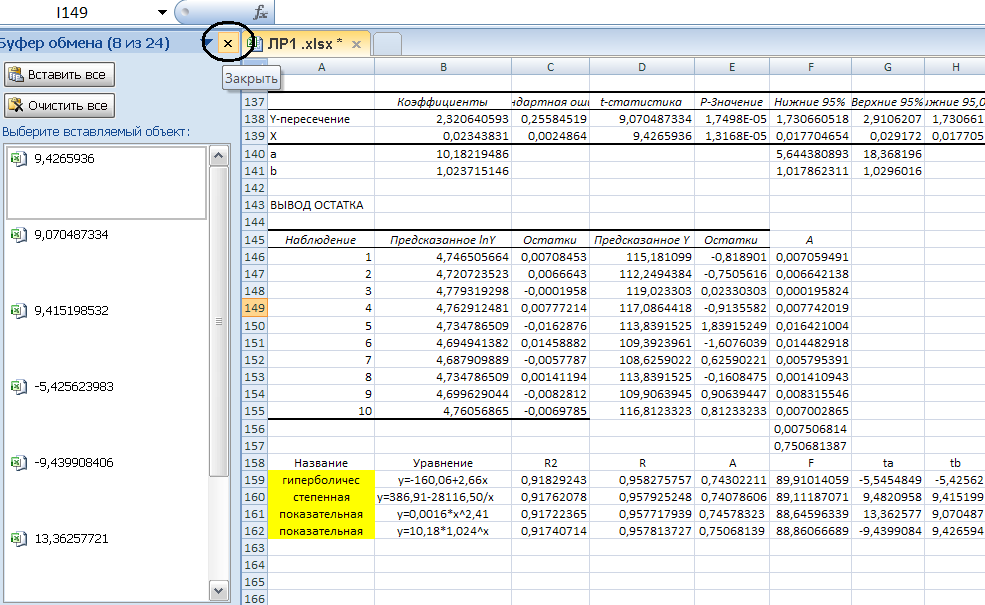
Получим таблицу



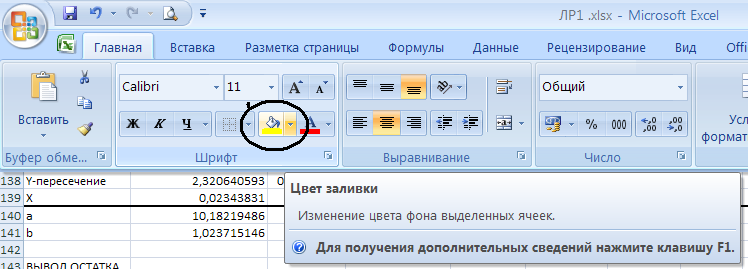
Очистим буфер обмена, нажав кнопку «Очистить все». Копируем в буфер обмена значения t-статистик.



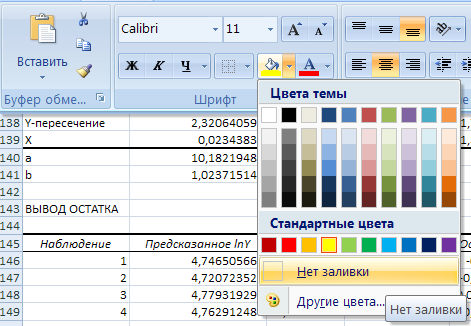
Вставим значения t-статистик в таблицу и закроем буфер обмена с помощью кнопки .



Чтобы убрать желтую заливку, выделим желтый участок на листе EXcel и на вкладке «Главная» в меню кнопки «Цвет заливки»

****

Щелкнем «Нет заливки»



Итак, имеем таблицу.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| название | уравнение | R2 | R | A | F |  |  |
| линейная |  | 0,91829 | 0,9583 | 0,7430 | 89,91014 | -5,54548 | 9,482096 |
| гиперболи-ческая |  | 0,91762 | 0,9579 | 0,7408 | 89,11187 | 13,36258 | -9,43991 |
| степенная |  | 0,91722 | 0,9577 | 0,7458 | 88,64596 | -5,42562 | 9,415199 |
| показательная |  | 0,91741 | 0,9578 | 0,7507 | 88,86067 | 9,070487 | 9,426594 |

**Выбор лучшей модели**

В каждом из столбцов таблицы R2, R, F, ,  отметим желтым цветом наибольшие результат, в столбце  – наименьший. Лучшей моделью является модель, у которой больше всего желтых клеток. Однако линейная модель имеет большое преимущество перед всеми остальными – простота в использовании, поэтому если ее характеристики не сильно проигрывают по сравнению с другими моделями, то выбирают линейную модель. Итак, y=-1.8056+0.34526.