**Факторы, влияющие на численность населения в России.**

Посмотрев интереснейшую лекцию Сергея Переслегина по демографии <https://www.youtube.com/watch?v=iZrqKvjiVHk> , я решил провести маленький эксперимент. Построить математическую модель факторов, оказывающих или не оказывающих слияние на численность населения. И вот, что получилось…

**Линейная регрессия (джан)**

Это наиболее распространенный способ показать зависимость какой-то переменной от других.

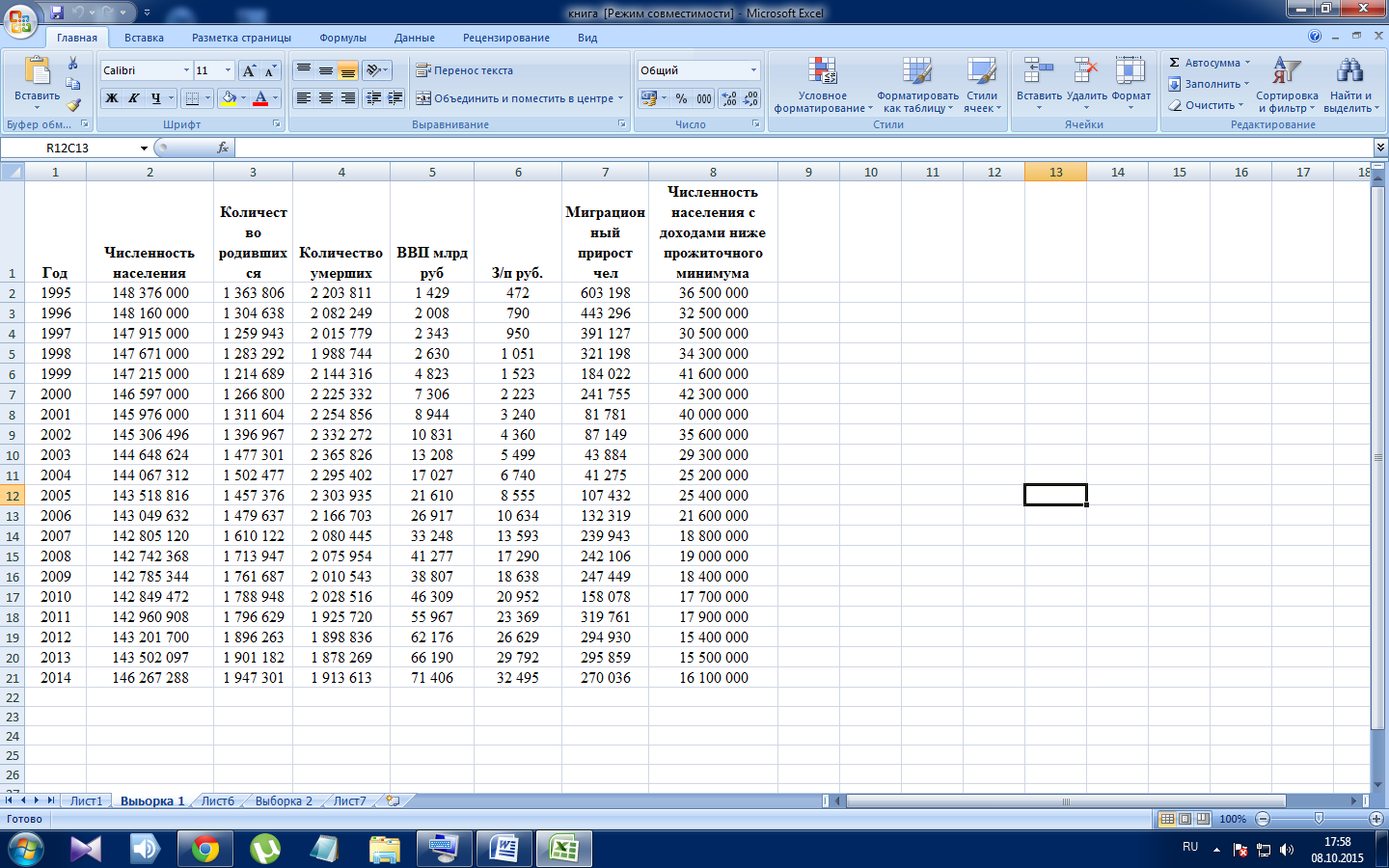
Общий вид модели линейной регрессии:

***Y=a0+a1x1+...+akxk***

где ***a*** — параметры (коэффициенты) регрессии, ***x*** — влияющие факторы, ***k*** — количество факторов модели.

### Исходные данные

Среди исходных данных нам необходим некий набор данных, который бы представлял из себя несколько последовательных или связанных между собой величин итогового параметра Y (численность населения) и такое же количество величин показателей, влияние которых мы изучаем (количество родившихся, количество умерших, ВВП, средняя з/п, миграционный прирост, численность населения с доходами ниже прожиточного минимума).



На рисунке выше показана таблица с этими самыми исходными данными, в качестве Y выступает показатель численность населения, а количество родившихся, количество умерших, ВВП, средняя з/п, миграционный прирост, численность населения с доходами ниже прожиточного минимума - это влияющие факторы, то бишь иксы (Х1,Х2, Х3, Х4, Х5, Х6).

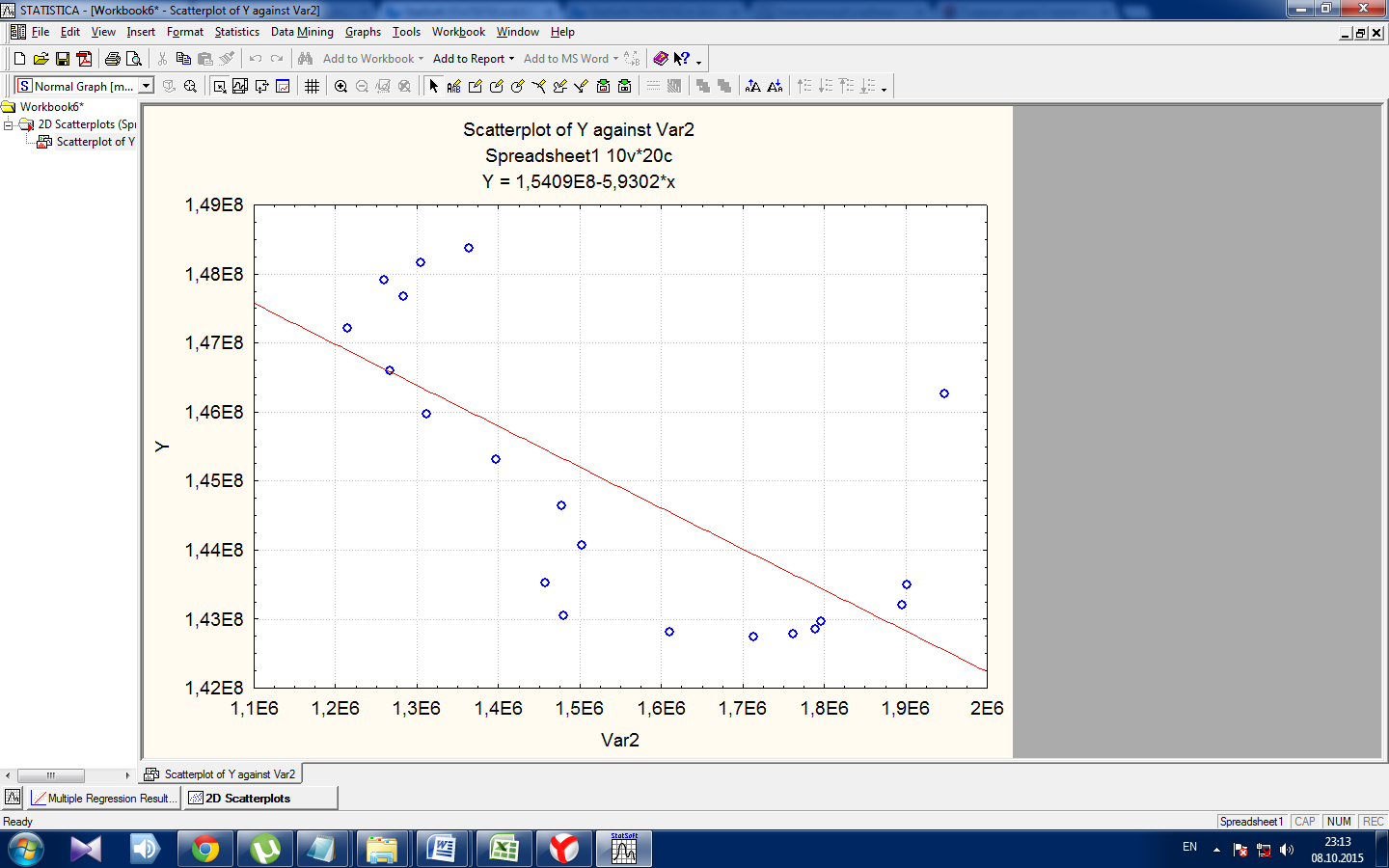
Сначала, я побаловался в программе «Статистика» с т.н. «Диаграммами рассеивания».

Диаграммы рассеяния

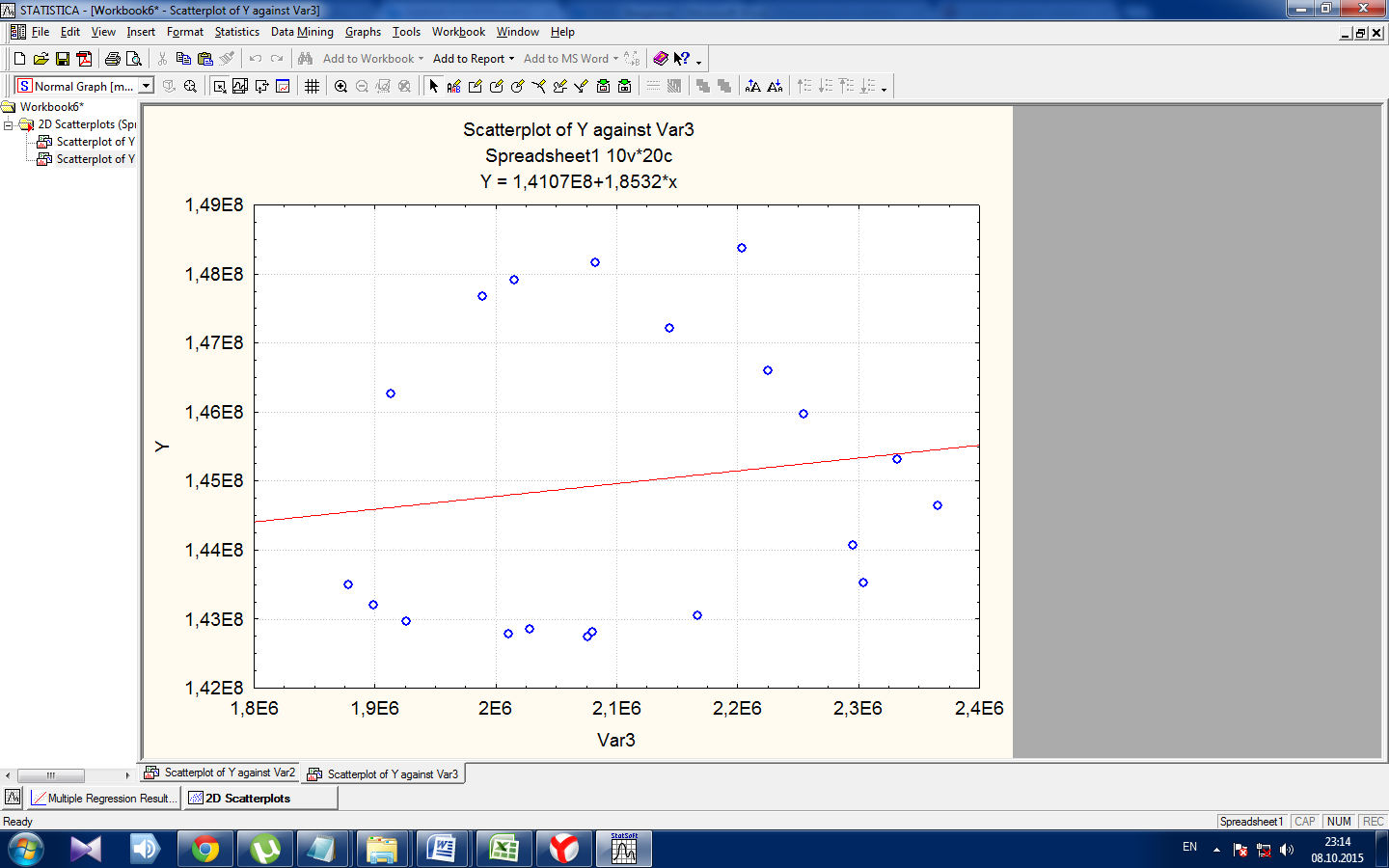
Двухмерные диаграммы рассеяния используются для **визуального исследования** зависимости между двумя переменными X и Y.

Данные изображаются точками в двухмерном пространстве. Две координаты (X и Y), которые определяют положение каждой точки, соответствуют значениям двух переменных. Если переменные сильно связаны, то множество точек данных принимает определенную форму (например, ложится на прямую линию или кривую, задаваемую определенным уравнением).

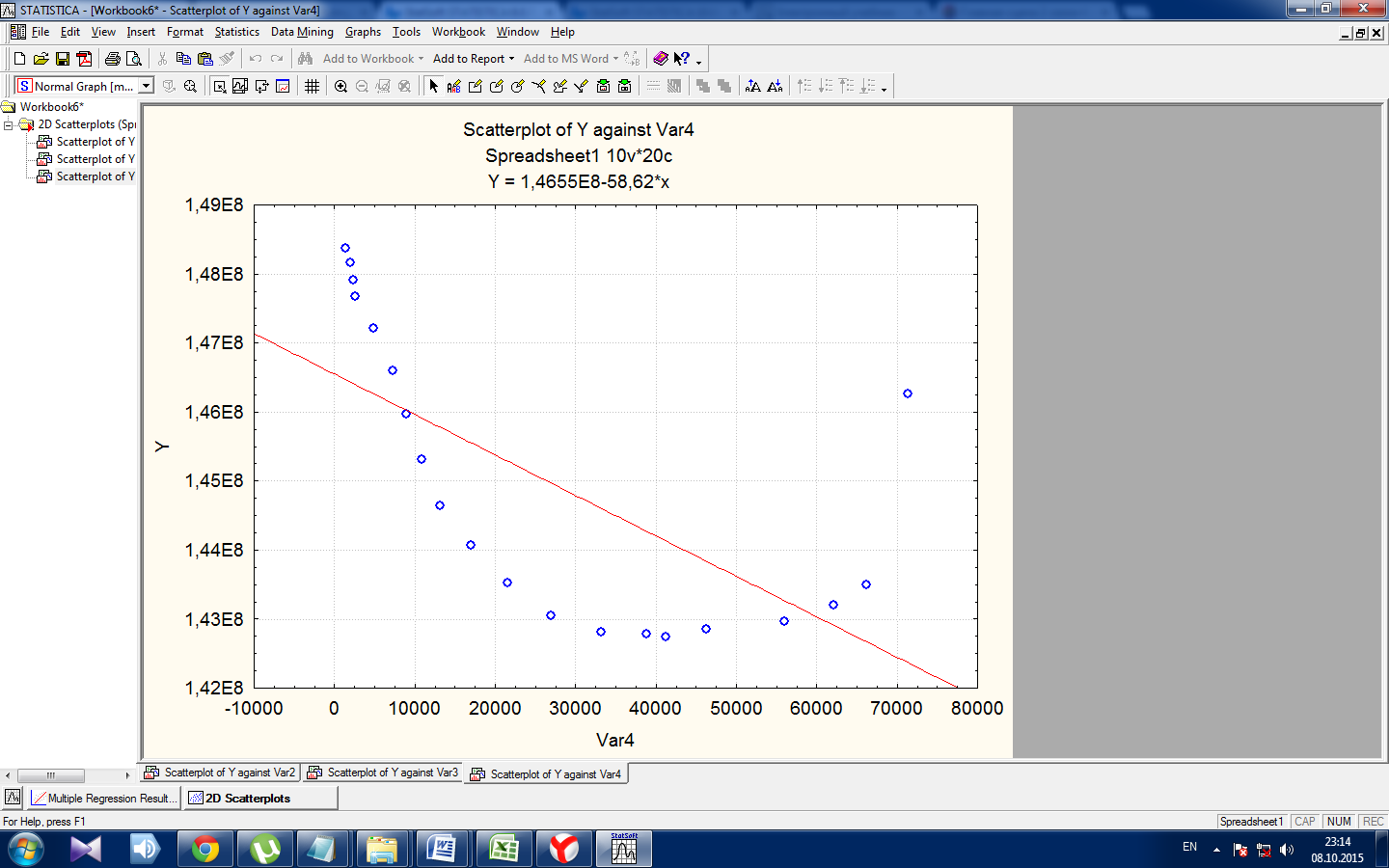
**Диаграмма 1. Зависимость численности населения от рождаемости**



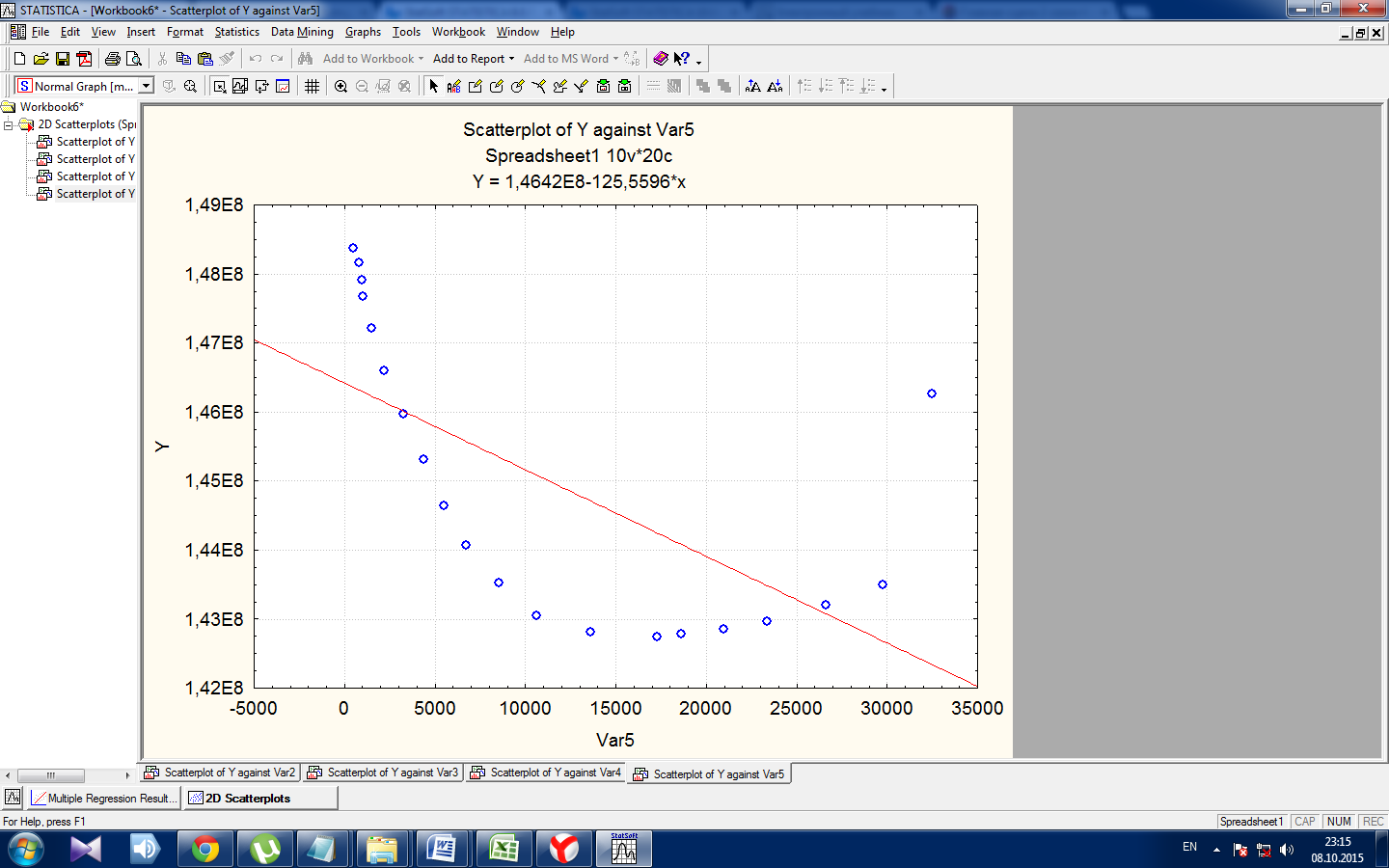
**Диаграмма 2. Зависимость численности населения от смертности**



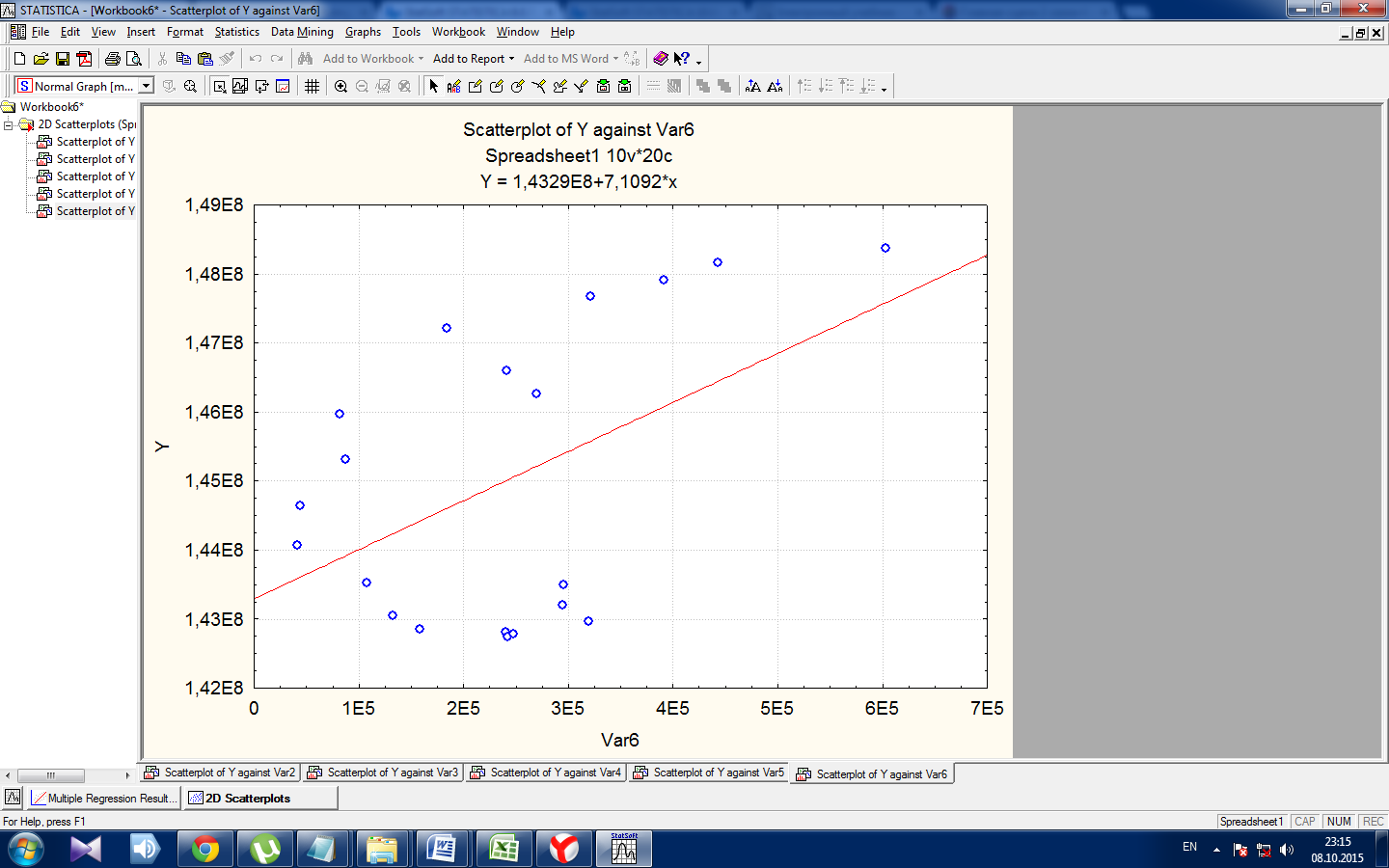
**Диаграмма 3. Зависимость численности населения от ВВП**



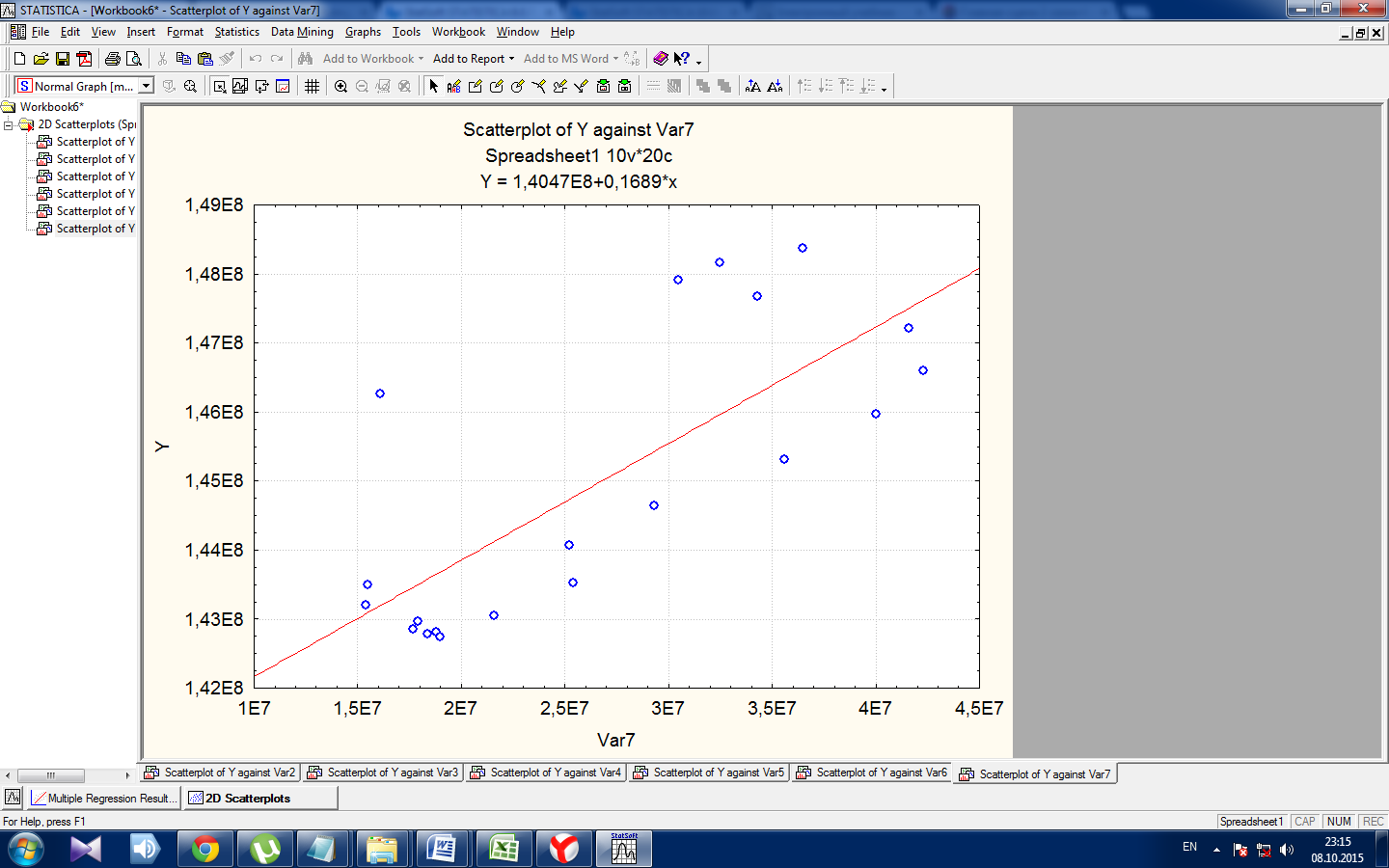
**Диаграмма 4. Зависимость численности населения от средней з/п**



**Диаграмма 5. Зависимость численности населения от миграционного прироста**



**Диаграмма 6. Зависимость численности населения от численности населения с доходом ниже прожиточного минимума**



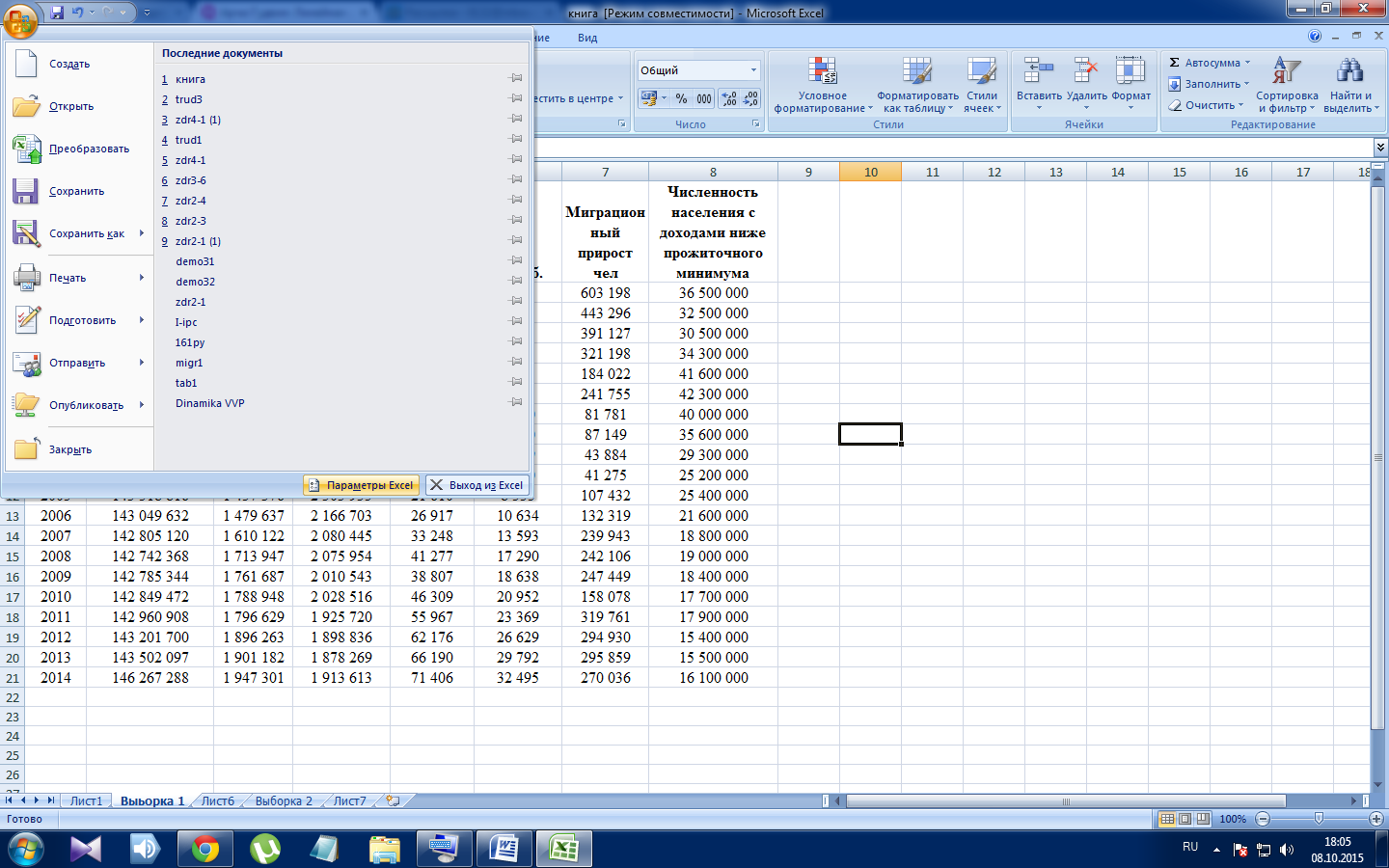
### В помощь тем, кто захочет войтиТь в этот прекрасный мир аналитики…..

### Пакет анализа

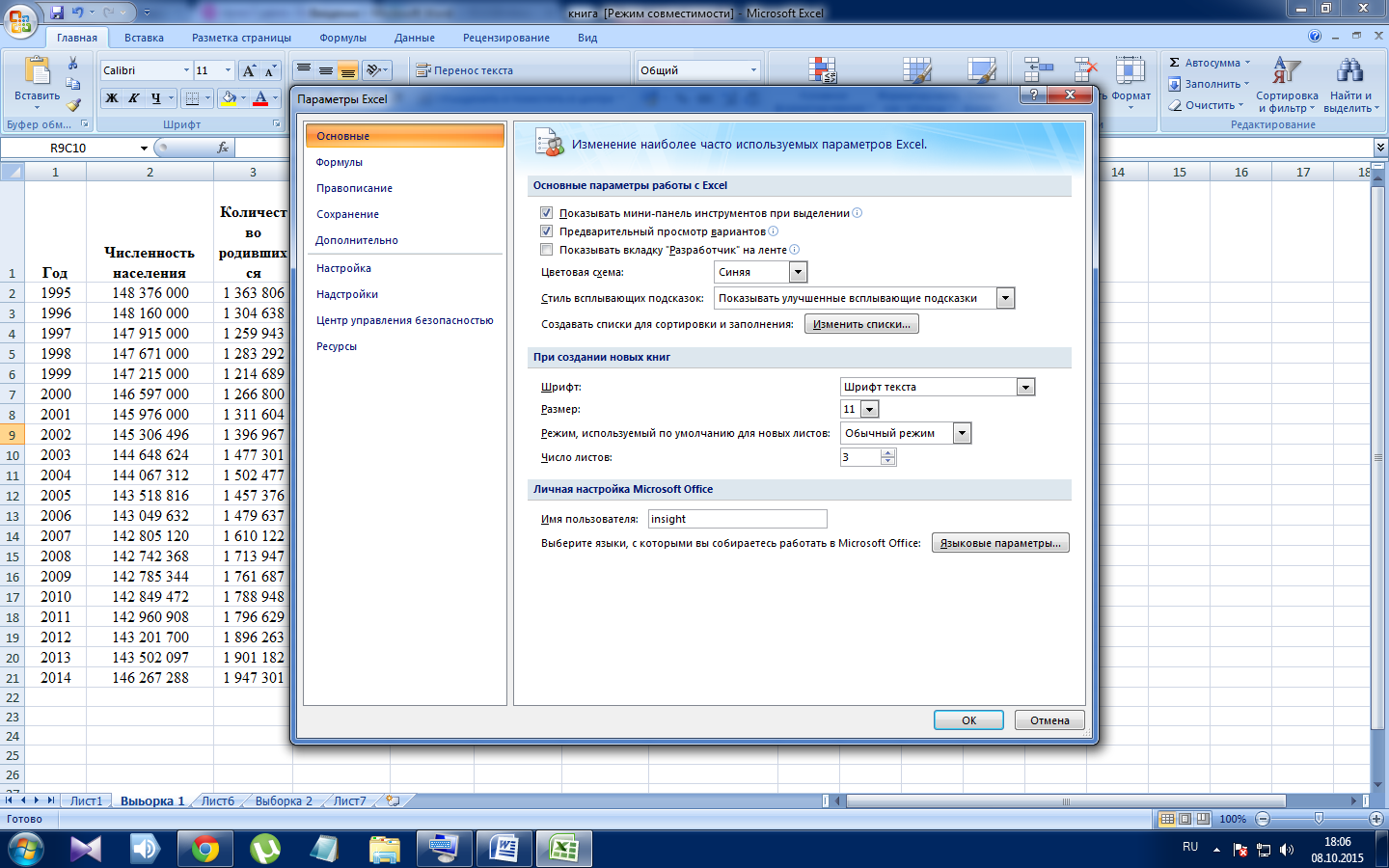
В Excel есть надстройка **Пакет анализа**, который является довольно мощным инструментом в помощь аналитику. Этот инструментарий, помимо всего прочего, умеет рассчитывать параметры регрессии.

### Активируем «Пакет анализа»

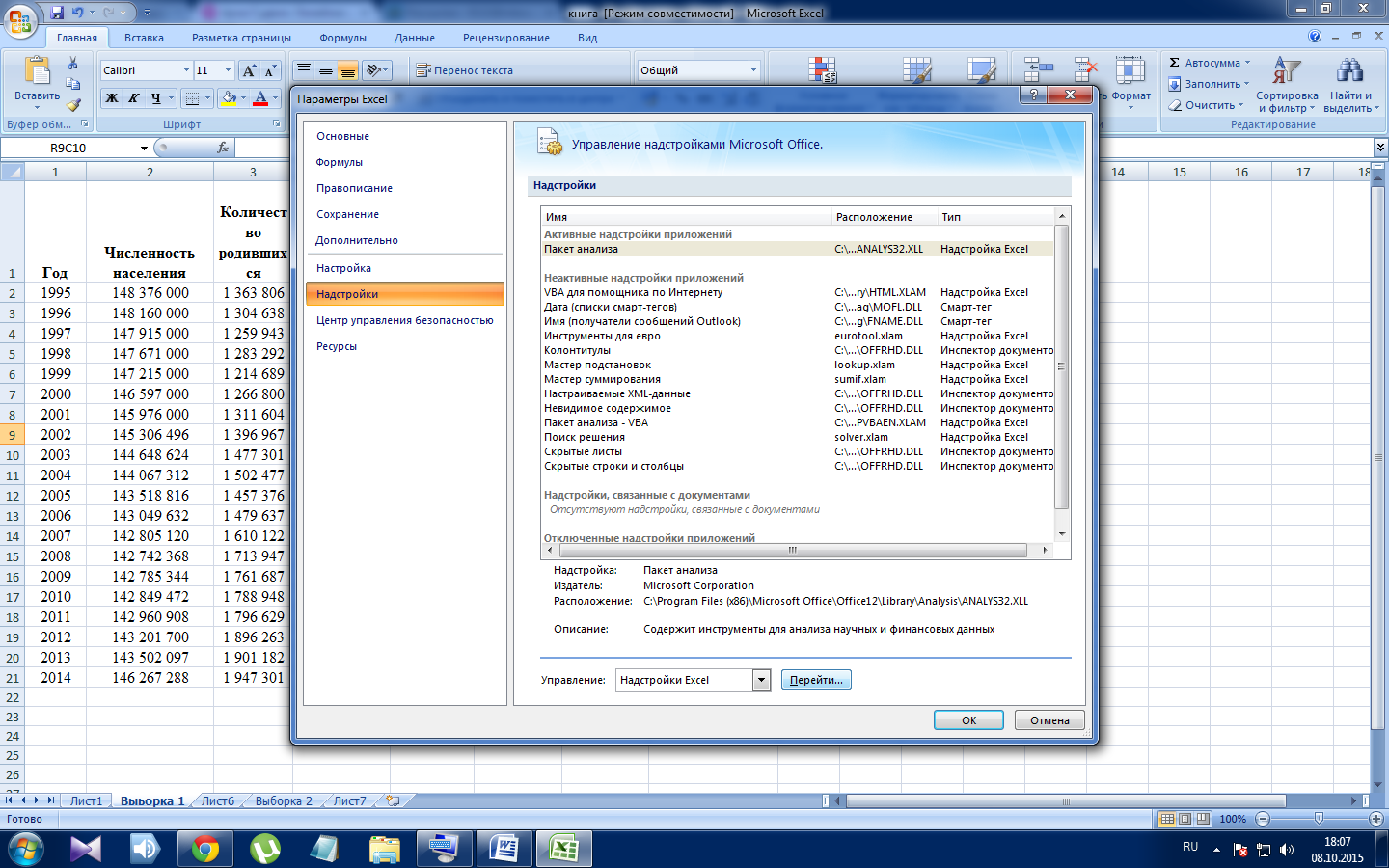
По умолчанию эта надстройка отключена и в меню вкладок вы ее не найдете, поэтому пошагово рассмотрим как ее активировать.

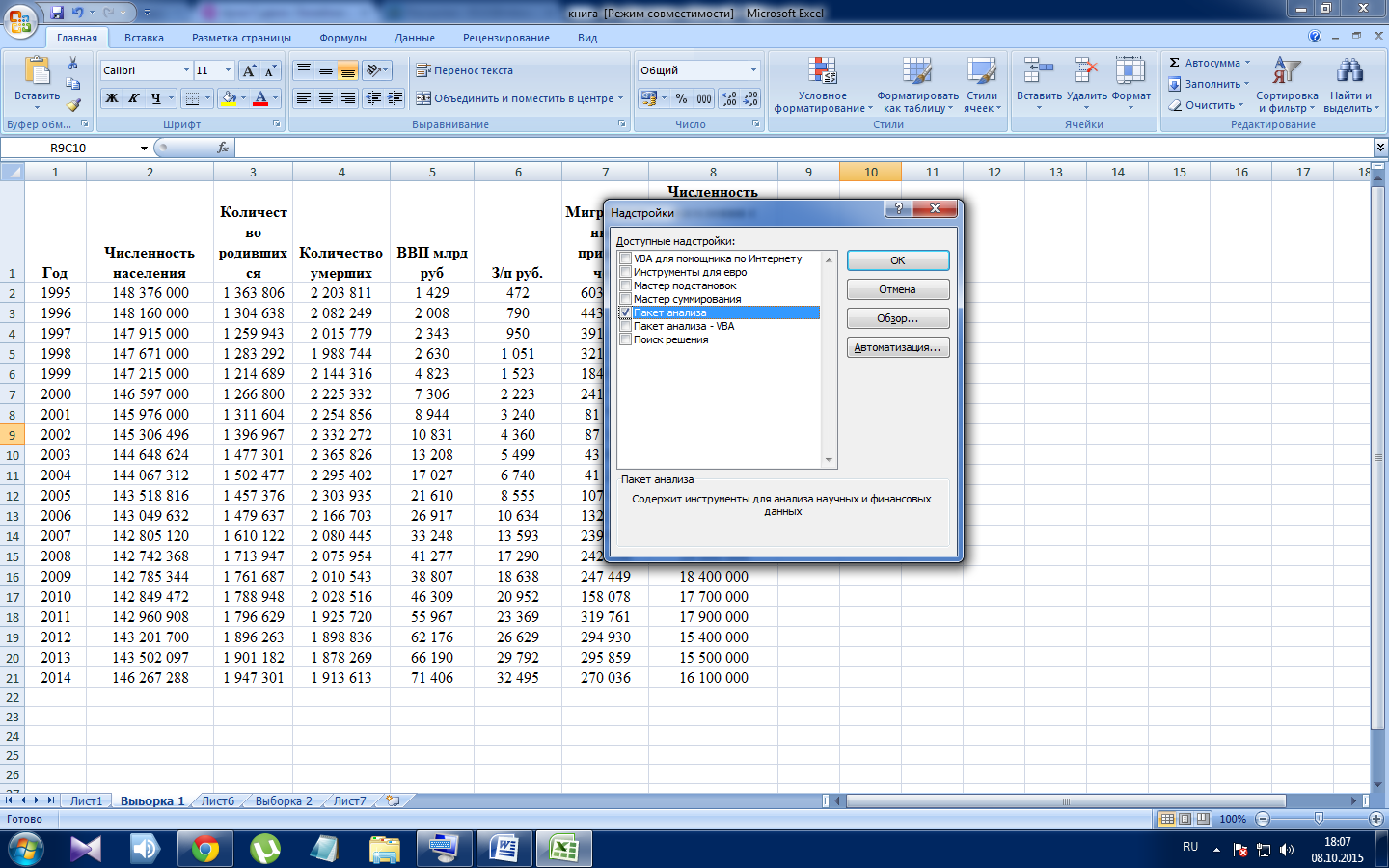


В эксель, слева вверху, активируем вкладку **Файл**, в открывшемся меню ищем пункт **Параметры**и кликаем на него.



В открывшемся окне, слева, ищем пункт **Надстройки**и активируем его, в этой вкладке внизу будет выпадающий список управления, где по умолчанию будет написано **Надстройки Excel**, справа от выпадающего списка будет кнопка **Перейти**, на нее и нужно нажать.

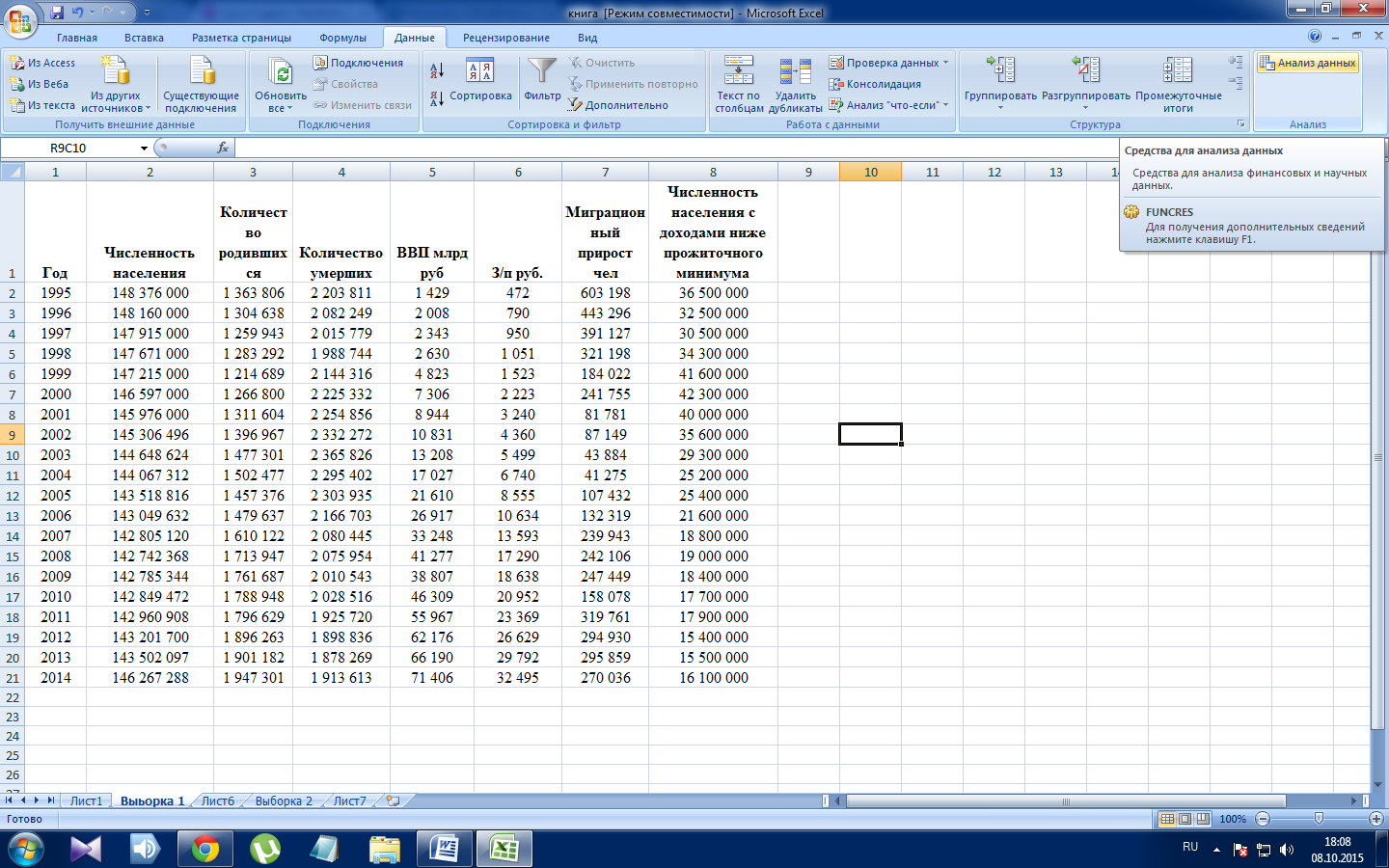




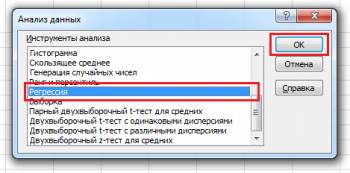
Всплывающее окошко предложит выбрать доступные надстройки, в нем необходимо поставить галочку напротив **Пакет анализа** и заодно, на всякий случай, **Поиск решения** (тоже полезная штука), а затем подтвердить выбор, кликнув по кнопочке **ОК**.

### Инструкция по поиску параметров линейной регрессии с помощью Пакета анализа

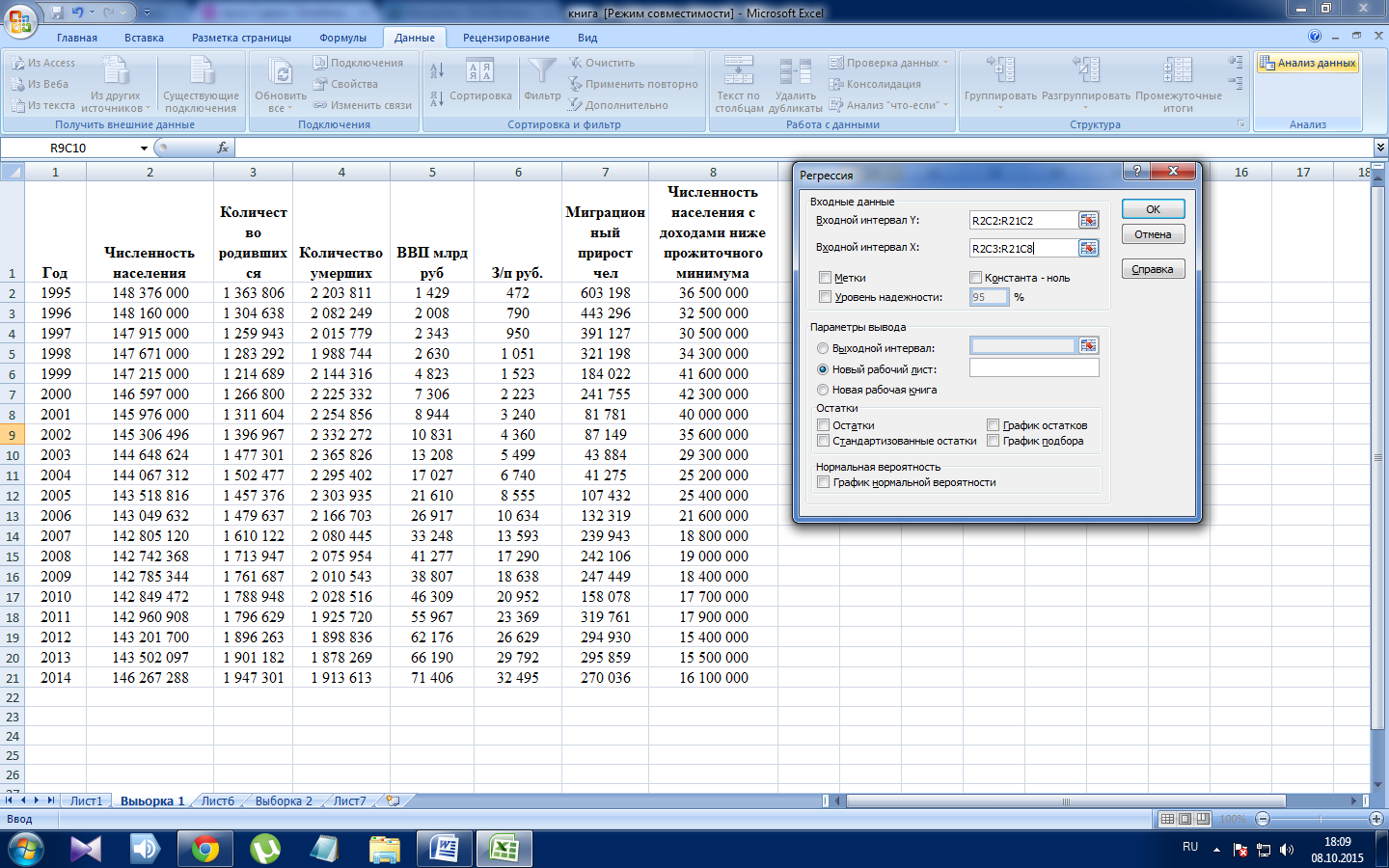
После активации надстройки Пакета анализа она будет всегда доступна во вкладке главного меню **Данные** под ссылкой **Анализ данных**



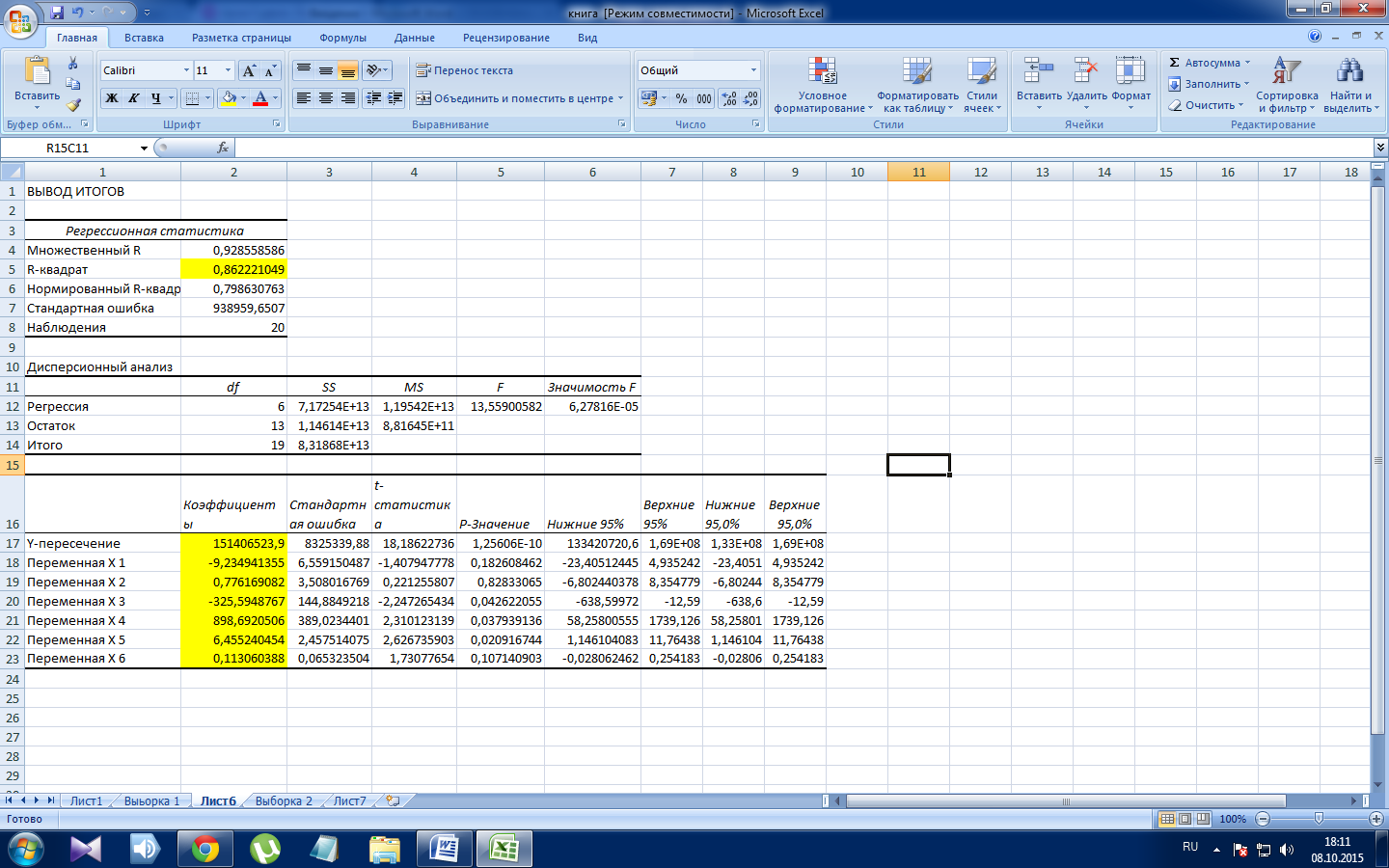
В активном окошке инструмента Анализа данных из списка возможностей ищем и выбираем Регрессия

[](http://archie-goodwin.net/_ld/3/01911606.png)

Далее откроется окошко для настройки и выбора исходных данных для вычисления параметров регрессионной модели. Здесь нужно указать интервалы исходных данных, а именно описываемого параметра (Y) и влияющих на него факторов (Х), как это на рисунке ниже, остальные параметры, в принципе, необязательны к настройке.

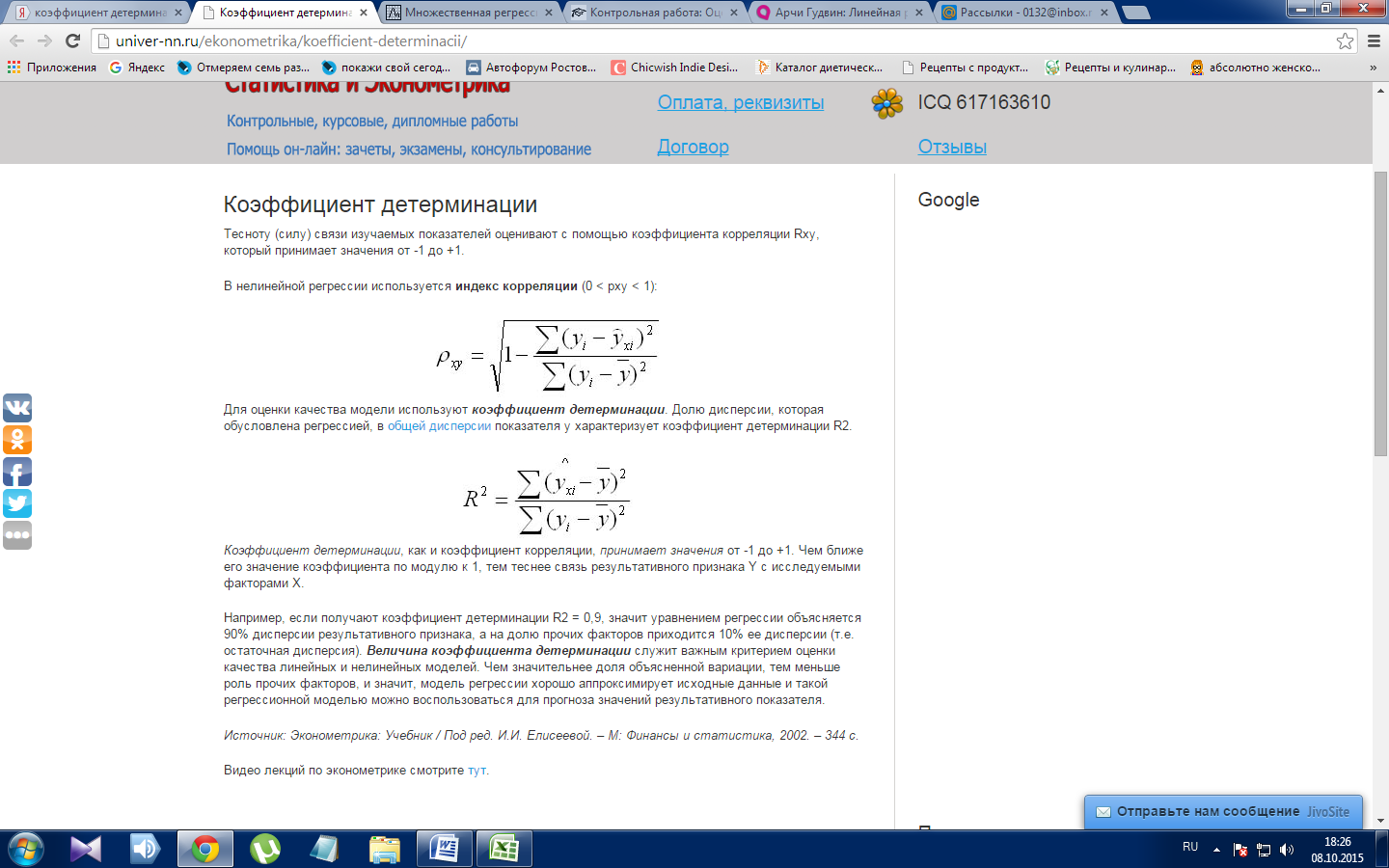


После того как выбрали исходные данные и нажали кнопочку ОК, Excel выдает расчеты на новом листе активной книги (если в настройках не было выставлено иначе), эти расчеты имеют следующий вид:



Ключевые ячейки залила желтым цветом именно на них нужно обращать внимание в первую очередь, остальные параметры значимость также немаловажны, но их детальный разбор требует пожалуй отдельного поста.

Итак, **0,8622** - это **R2** - коэффициент детерминации, показывающий что на 86,22% расчетные параметры модели, то есть сама модель, объясняют зависимость и изменения изучаемого параметра - **Y** от исследуемых факторов - **иксов**. Если утрировано, то это показатель качества модели и чем он выше, тем лучше. Понятное дело, что он не может быть больше 1 и считается неплохо, когда R2 выше 0,8, а если меньше 0,5, то резонность такой модели можно смело ставить под большой вопрос.



Теперь перейдем к коэффициентам модели:

**151 406 523,9**- это **a0** - коэффициент который показывает какой будет Y в случае, если все используемые в модели факторы будут равны 0, подразумевается что это зависимость от других неописанных в модели факторов;   
**-9,23**- **a1** - коэффициент, который показывает весомость влияния фактора x1 на Y, то есть количество родившихся в пределах данной модели влияет на показатель численность населения. Знак минус показывает что это влияние отрицательно, то есть чем больше предприятий, тем меньше экономически активного населения, как бы это ни было парадоксальным по смыслу;   
**0,78**- **a2** - коэффициент влияния количества умерших на величину численность населения;   
**-325,59**- **a3**- коэффициент влияния ВВП на величину численность населения

**898,69**– **a4**- коэффициент влияния средней з/п на величину численность населения,

**6,46**– **a5**- коэффициент влияния прироста миграции на величину численность населения.

**0,11**– **a6**- коэффициент влияния численности населения с доходами ниже прожиточного минимума на величину численность населения,

Соберем рассчитанные коэффициенты в модель:

***Y = 151 406 523,9 - 9,23x1 + 0,78x2 – 325,59x3+ 898,69x4+ 6,46x5+ 0,11x6***

Собственно, это и есть линейная регрессионная модель, которая для исходных данных, используемых в примере, выглядит именно так.

ПРОДОЛЖЕНИЕ В ПРОГРАММЕ СТАТИСТИКА!

Однако среди переменных были незначимые, которые портили качество модели.

### В Statistica я исключил незначимые переменные и остались только 2 переменные: «миграционный прирост» и «численность населения с доходом ниже прожит минимума».

### 

### Здесь коэффициенты R должны быть близки к 1, больше 0,7 – уже очень хорошо и F должен быть, чем выше, тем лучше. Он по формуле вычисляется, но в данном случае значение 32,00760 – это хорошо.

### Таким образом, путем откидывания ненужных переменных, оказывающих незначительное влияние на У получилась модель, вполне приличная.

### КАК Я ВЫКИДЫВАЛ ПЕРЕМЕННЫЕ ПОКАЗАНО ПОДРОБНО В МОДЕЛИ 2 НИЖЕ

### Модель 2

### 

### Здесь больше факторов, добавились браки, разводы, заболеваемость, численность безработных, численность инвалидов. Однако сократились количество наблюдений: с 2000 года по 2013 г.

### 

### Скрин из Statistica (те же показатели, в принципе)

### 

### Основная проблема применения (выборочного) R2   заключается в том, что его значение увеличивается (*не* уменьшается) от добавления в модель новых переменных, даже если эти переменные никакого отношения к объясняемой переменной не имеют. Поэтому сравнение моделей с разным количеством признаков с помощью коэффициента детерминации, вообще говоря, некорректно. Для этих целей можно использовать альтернативные показатели.

### Переменные (обвел в красный прямоугольник) все синего цвета. Значимые переменные должны быть красного цвета. Тут как раз та ситуация, что переменных много и коэффициент показывающий качество модели R2  очень близок к 1 только поэтому, а не, потому что модель действительно хорошая.

### Исключаем переменные по очереди.

### 

### Опираемся на t(2) показатель должен быть, чем выше, тем лучше и на p-level – должен быть максимально близко к нулю.

### Самое маленькое значение t(2) у переменной «родивш» = 0,37382. Исключаем ее и смотрим что получилось:

### 

### Переменная «ниже прож мин» стала красного цвета – стала значимой. Исключаем переменные дальше:

### 

### Самое маленькое значение t(3) у переменной «браки» = (-0,73637). Исключаем ее и смотрим что получилось:

### 

### Появилось сразу еще 4 значимых переменных (все что красным отмечены). Получается, что исключенные нами переменные просто скрывали значимость действительно нужных модели переменных. Так же отметим, что F показатель растет. Чем он выше, тем лучше качество модели.

### Продолжим исключать:

### 

### Самое маленькое значение t(4) у переменной «умерш» = (1,19610). Исключаем ее и смотрим что получилось:

### 

### Далее

### 

### Исключаем переменную «з/п».

### 

### Исключим последнюю незначимую переменную «ВВП»

### 

### Все переменные красного цвета. Все переменные значимы. Коэффициенты прочности модели высокие.

### 

### В овале отмечены коэффициенты регрессии. Обратим внимание, что некоторые коэффициенты со знаком минус. например «миграц», «разводы», «безработ», «инвал». Этого говорит об обратном влиянии на У (численность населения). Другими словами: чем ниже миграция – тем выше численность, чем меньше разводов – тем выше численность населения ит.д.

### Если опираться на то что чем ниже значение t (кружком отметил) тем хуже, соответственно чем выше – тем больше влияние на У (численность населения). Наивысшее влияние оказывает переменная «ниже прож мин». Все остальные примерно в равных степенях.

### Выводы получились, скорее… прогнозируемые, но интересные! ЖАЛЬ, ЧТО ДАННАЯ МОДЕЛЬ НЕ УЧИТЫВАЕТ И ДРУГИЕ, ЛАТЕНТНЫЕ ФАКТОРЫ, КОТОРЫЕ НЕ УЧТЕНЫ В МОДЕЛИ.

### Ещё, хочется добавить, что данная модель, вряд ли подойдет для стран Ближнего Востока, да и ряда других стран. Скорее всего, там всё по-другому. И, влияющие факторы, несколько иные.

### Автор благодарит Анастасию В. Забелину, за неоценимую помощь в проведении этого маленького исследования.

### © Аллавердян В.В., 2017 г.

### 