МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ  
федеральное государственное АВТОНОМНОЕ образовательное учреждение высшего профессионального образования  
«Национальный исследовательский ядерный университет «МИФИ»  
  
  
**Димитровградский инженерно-технологический институт –**филиал федерального государственного автономного образовательного учреждения высшего  
профессионального образования «Национальный исследовательский ядерный университет «МИФИ»  
**(ДИТИ НИЯУ МИФИ)**

**Лабораторная работа №4  
по курсу «Основы математической статистики и   
планирование эксперимента»**

Составил: доцент кафедры   
 высшей математики   
 канд. экон. наук   
 Кожухова В.Н.

Димитровград 2017

**Лабораторная работа №4  
по курсу «Основы математической статистики и   
планирование эксперимента»**

**Парная линейная регрессия**

***Содержание***

[Цель работы 3](#_Toc338149361)

[Задание 3](#_Toc338149362)

[Теоретическая часть 4](#_Toc338149363)

[Эконометрические модели 4](#_Toc338149364)

[Этапы построения эконометрической модели 4](#_Toc338149365)

[Метод наименьших квадратов (МНК, method of Least Squares, LS) 5](#_Toc338149366)

[Нелинейная регрессия 8](#_Toc338149367)

[Проверка качества уравнения регрессии 10](#_Toc338149368)

[Критерии общей точности модели 11](#_Toc338149369)

[Критерии выбора модели 14](#_Toc338149370)

[Проверка качества оценок параметров 14](#_Toc338149371)

[Пример выполнения работы 16](#_Toc338149372)

[1. Визуальный анализ исходных данных. 17](#_Toc338149373)

[2. Корреляционный анализ. 17](#_Toc338149374)

[3. Идентификация моделей. 18](#_Toc338149375)

[4. Оценка общего качества моделей. 21](#_Toc338149376)

[5. Выбор наилучшей модели. 24](#_Toc338149377)

[6. Оценка качества идентификации параметров. 27](#_Toc338149378)

[7. Прогнозирование по модели. 28](#_Toc338149379)

[Оформление отчета 30](#_Toc338149380)

[Контрольные вопросы 30](#_Toc338149381)

[Список рекомендуемой литературы 31](#_Toc338149382)

## Цель работы

Научиться оценивать значения параметров линейной и нелинейной, но сводимой к линейной парной регрессии; оценивать качество полученной регрессии и параметров; сравнивать модели между собой и осуществлять их выбор.

## Задание

Имеются две выборки равного объема для показателей *Y* и *X*. Предполагается наличие зависимости уровней *Y* от *X*.

1. Визуальный анализ исходных данных.

Построить график зависимости *Y* от *X*. Выдвинуть предположение о наличии и функциональном виде зависимости.

1. Корреляционный анализ.

Проверить наличие корреляционной зависимости между *Y* и *X*, *Y* и *X*2, *Y* и 1/*X*, Y и ln*X*, ln*Y* и *X*, ln*Y* и ln*X*.

1. Идентификация моделей.

Оценить параметры традиционных моделей регрессии:

* 1. линейной;
  2. параболической;
  3. логарифмической;
  4. обратной (гиперболической);
  5. показательной (экспоненциальной);
  6. степенной.

1. Проверка общего качества моделей.

Проверить общее качество полученных уравнений регрессии. Сравнить с результатами, полученными в п.2. Отсеять неудовлетворительные модели.

1. Выбор наилучшей модели.

Из оставшихся моделей выбрать наилучшую, учитывая количественные критерии и качественный анализ.

1. Оценка качества идентификации параметров.

Проверить статистическую значимость оценок параметров для выбранной модели и построить их доверительные интервалы.

1. Прогнозирование по модели.

Построить точечную и интервальную оценку *Y* при заданном значении *X = x*.

Если в п.4 были отсеяны все модели, п.6-7 рассчитать для линейной модели. Если в п.5 была выбрана параболическая модель, п.6 рассчитать для линейной модели.

## Теоретическая часть

### Модели

Модели **пространственной** динамики описывают взаимосвязи между различными показателями.

Модели **временнóй** динамики описывают развитие одного фактора во времени.

**Смешанные** – пространственно-временные модели, описывают взаимодействие нескольких показателей с учетом их развития во времени.

Простейшая пространственная модель – модель **парной линейной регрессии**:

,

где  – зависимая переменная (регрессор),  – независимая переменная (фактор),  – параметры модели,  – стохастическая компонента (случайные остатки, ошибка, погрешность);  – номер наблюдения;  – общее число наблюдений.

Это *теоретическое* уравнение регрессии, для которого необходимо оценить значения параметров по конкретным выборкам X и Y. В результате получают *эмпирическое* уравнение регрессии, модельные (ожидаемые) значения  и оценки ошибок :

.

### Этапы построения модели

|  |  |
| --- | --- |
| 1. Постановка задачи, анализ предметной области – выбор исследуемых показателей |  |
| 1. Сбор исходных данных – формирование выборок (рядов), по которым будет строиться модель |  |
| 1. *Спецификация* модели (структурная идентификация) – определение общего вида модели |  |
| 1. *Идентификация* модели (параметрическая идентификация) – оценка значений параметров модели |  |
| 1. *Верификация* модели – проверка качества модели и ее параметров |  |
| 1. *Интерпретация* модели – анализ результатов моделирования |  |

Все этапы взаимосвязаны между собой, зачастую после выполнения какого-то этапа приходится вернуться к предыдущим. Например, после спецификации модели могут потребоваться дополнительные статистические данные, или после верификации модели она оказывается неудовлетворительной и приходится рассматривать другую модель.

Как правило, выдвигается несколько гипотез о возможном виде модели, из которых затем выбирается наилучшая. Критерии выбора могут быть различными.

### Метод наименьших квадратов (МНК, method of Least Squares, LS)

*Суть метода* заключается в минимизации отклонений модели от исходных данных по всем точкам выборки. Чтобы отклонения модели с противоположными знаками не компенсировали друг друга при суммировании, они возводятся в квадрат.

Общий вид:



или



или

.

Если параметры  входят в модель линейно, то существует единственный минимум. Чтобы его найти, необходимо приравнять частные производные по всем параметрам нулю:



Для парной линейной регрессии:















Таким образом:

.

 – [выборочный] коэффициент регрессии *Y* по *X*, показывает на сколько единиц изменится *Y* при изменении *X* на 1.

 и  – оценки, зависящие от конкретной выборки, т.е. *случайные величины.*

*Оптимальными* считаются *состоятельные*, *несмещенные* и *эффективные* оценки (BLUE-оценки – best linear unbiased estimates).

**Состоятельность** – при увеличении объема выборки оценка приближается к своему истинному значению (сходится по вероятности):

.

При объеме выборки , стремящемуся к бесконечности, вероятность  того, что отклонение оценки  от истинного значения  будет меньше некоторого малого положительного числа , стремится к 1.

**Несмещенность** – среднее значение оценки равно ее истинному значению:

.

Иногда говорят об асимптотической несмещенности:

,

т.е. несмещенность достигается при увеличении объема выборки.

**Эффективность** – оценка обладает наименьшей дисперсией (среди всех методов):

.

Поскольку все методы учесть невозможно, говорят об эффективности оценок в некотором классе методов.

**Теорема Гаусса-Маркова:**

Оценки параметров линейной регрессии являются эффективными в классе линейных несмещенных оценок, если выполняются следующие условия (**условия Гаусса-Маркова**):

1. Математическое ожидание стохастической компоненты  равно нулю:

.

Если это условие не выполняется, то модель содержит систематическую ошибку. Когда в модели присутствует свободный член (константа *b*0), по результатам МНК всегда выполняется . Если в исходных данных присутствовала систематическая ошибка, то она просто прибавится к *b*0, поэтому обычно считают, что данное условие выполняется автоматически. Проверить его выполнение практически невозможно.

1. Дисперсия  постоянна для всех наблюдений (*гомоскедастичность*):

.

Нарушение этого условия называется *гетероскедастичностью* и нередко возникает на практике.

Существуют специальные тесты для обнаружения гетероскедастичности (Голдфелда-Квандта, Глейзера, ранговой корреляции Спирмена и др.). Вместо МНК при этом применяется МВНК – метод взвешенных наименьших квадратов.

1. Случайные отклонения для различных наблюдений  и  являются некоррелированными (*отсутствие автокорреляции*):

.

Это условие также зачастую нарушается на практике, особенно для временных моделей.

Для обнаружения автокорреляции используется тест Дарбина-Уотсона, для идентификации таких моделей – обобщенный метод наименьших квадратов (ОМНК). Если выполняется условия 3 и 5, то  и  являются независимыми.

1. Случайные отклонения  должны быть некоррелированы с объясняющей переменной *X*.

.

Нарушение данного условия означает, что в случайных остатках содержится один или несколько неучтенных факторов. Решением в этом случае может быть переход от линейной модели к нелинейной или от парной регрессии к множественной. Также возможно, что неправильно определены зависимая и независимая переменные.

Однако, как и для условия 1, обнаружить нарушение этого условия нельзя. Если в  содержится линейная зависимость от , то она будет включена в *b*1, и для оценок  будет выполняться всегда.

Для временных моделей данное условие выполняется автоматически, поскольку в них в качестве независимой переменной выступает время, неслучайная величина, которая не может коррелировать со случайной.

1. Стохастическая компонента  имеет нормальное распределение:

.

Данное условие не является обязательным и необходимо лишь для проверки качества модели и оценок параметров. На малых выборках проверить его выполнение практически невозможно. Иногда его заменяют более мягким требованием симметричности закона распределения. Обычно считают, что оно выполняется по закону больших чисел, поскольку  аккумулирует все множество факторов, не учтенных в модели.

При выполнении всех пяти условий Гаусса-Маркова модель парной линейной регрессии называется *классической нормальной линейной регрессионной моделью*.

### Нелинейная регрессия

Нелинейные модели регрессии позволяют описывать более сложную форму зависимости *Y* от *X*.

Различают *нелинейные по переменным* и *нелинейные по параметрам* модели.

Наиболее распространенные нелинейные по переменным, но линейные по параметрам модели:

* параболическая ;
* полиномиальная ;
* логарифмическая ;
* обратная (гиперболическая) .

Такие модели могут быть напрямую идентифицированы с помощью МНК. Например, для гиперболической модели можно принять:

 – линейная модель.

Для параболической модели необходимо оценить значения трех параметров:









Данную систему проще решать в матричном виде:



По методу определителей:

,

где  – определитель матрицы *А*, , ,  – частные определители системы.

Модели, нелинейные по параметрам, зачастую могут быть линеаризованы с помощью различных преобразований.

Наиболее распространенные *нелинейные по параметрам* модели:

* показательная (экспоненциальная) :

;

* степенная :

;

где , , ;

Обратите внимание, что для того, чтобы реализация была возможна, стохастическая компонента  включается в модель **мультипликативно**. После линеаризации новая стохастическая компонента  входит в модель **аддитивно**.

Применяя МНК к линеаризованным моделям необходимо учитывать, что условия Гаусса-Маркова должны выполняться для , а не . Несмещенность, эффективность и состоятельность также гарантируется для , а не .

Тогда про  можно сказать, что она имеет несимметричное логнормальное распределение, , , автокорреляция и корреляция с  отсутствует.

Визуально модель будет выглядеть как гетероскедастическая, поскольку  – это относительная погрешность и условие  означает постоянство доли ошибки, а нее ее абсолютного значения.

0

X

0

X

Y

Y

F(ε)



Из-за этого мультипликативные и аддитивные модели трудно сравнивать между собой. Формально для мультипликативной модели можно вычислить погрешности , но характеристики их будут другими.

Если модель линеаризовать не удается, применяют нелинейный МНК (НМНК), т.е. численное решение задачи минимизации и системы нелинейных алгебраических уравнений. Для НМНК доказана лишь состоятельность получаемых оценок параметров при выполнении условий Гаусса-Маркова.

### Проверка качества уравнения регрессии

Как правило, выдвигается несколько гипотез о возможном виде модели, из которых затем выбирается наиболее адекватная.

**Адекватность** модели означает ее соответствие исходным данным с точки зрения цели моделирования.

Следует отличать *адекватность* модели от ее ***точности***, т.е. близости модельных данных к исходным. Адекватность является более общим понятием, включающим в себя точность. Кроме того, точность является количественной характеристикой, а адекватность – качественной.



Таким образом, проверка точности модели необходима, но не достаточна для обоснования ее адекватности. Дополнительно требуется качественное обоснование модели.

Проверка качества уравнения регрессии включает проверку ее общей точности, а также точности оценок параметров. Необходимо проверить а) является ли модель и ее параметры статистически значимыми (позволяет ли имеющаяся выборка судить о генеральной совокупности); б) какова практическая ценность модели (насколько информация, полученная по модели, снимает неопределенность в отношении регрессора Y).

### Критерии общей точности модели

Существует большое количество критериев точности, используемых как в экономике, так и в технике.

Наиболее простыми можно считать **средние абсолютное и относительное отклонения** (mean absolute error, mean absolute percentage error):

,

.

Первый критерий выражается в тех же единицах измерения, что и моделируемый ряд динамики, поэтому он может быть использован только для сравнения моделей одного и того же ряда динамики. С другой стороны, он позволяет оценить «физическое содержимое» погрешности моделирования.

Среднее относительное отклонение является безразмерной величиной и позволяет судить о точности модели и сравнивать их между собой. Однако ее значение также во многом зависит от уровней ряда динамики. Если значения  велики по сравнению со своим разбросом, то MAPE-оценка будет малой, а если близки к нулю, то MAPE-оценка будет большой вне зависимости от точности модели. Если же имеются наблюдения, строго равные нулю, то использовать относительные величины вообще невозможно.

Существует несколько менее распространенных, похожих на MAPE, оценок: симметричная MAPE – sMAPE, взвешенная MAPE – WAPE, средняя квадратичная ошибка – SME и корень из нее – RSME и др.

Чаще всего эти оценки используются при исследовании временной динамики.

**Коэффициент корреляции** (выборочный) определяет силу линейной зависимости между показателями:

,

.

Чем ближе модуль  к 1, тем точнее модель. При  между X и Y существует функциональная зависимость, при  линейная связь полностью отсутствует.

Знак  определяет направление зависимости (положительная или отрицательная).

Для нелинейных моделей коэффициент корреляции рассчитывается для их линеаризованной формы, например для логарифмической модели рассчитывается

.

Как и оценки параметров, выборочный коэффициент корреляции является случайным, зависящим от выборки. Поэтому его значение следует оценивать с позиций статистики.

Проверка *гипотезы* *о статистической значимости коэффициента корреляции*:

 – линейная зависимость отсутствует,

 – линейная зависимость присутствует.

Для проверки гипотезы стоится *t-статистика*:

,

которая сравнивается с критической точкой распределения Стьюдента , где  – уровень значимости,  – число степеней свободы, количество наблюдений минус число параметров.

Если , то нет оснований для отклонения .

Если , то  отклоняется в пользу .

Обычно считают, что если , то линейная зависимость отсутствует, если  – зависимость слабая, если  – присутствует сильная линейная зависимость.

**Коэффициент детерминации** (coefficientof determination) *R*2:

.

 (сумма квадратов отклонений, unexplained sum of squares) – мера остаточного, не объясненного моделью разброса исходных данных.

(общая сумма квадратов, total sum of squares) – мера общего рассеивания  относительно линии математического ожидания .

*Смысл* : какая доля зависимого показателя не является случайной, т.е. описывается моделью.

Другая интерпретация: насколько лучше найденная модель объясняет исследуемую зависимость, чем горизонтальная прямая .

Для линейных моделей . Чем ближе коэффициент детерминации к единице, тем точнее модель. При  модель проходит точно через исходные данные.

Для нелинейных моделей коэффициент детерминации может быть отрицательным , если модель совершенно не объясняет показатель (даже хуже, чем просто горизонтальная прямая). Причиной может быть, например, вычислительная ошибка, неверный выбор функционального вида модели.

Для линейных моделей можно использовать и другую формулу:

.

 (объясненная сумма квадратов, explained sum of squares) – мера объясненного моделью разброса.

Также для линейных моделей . Поэтому для нелинейных моделей иногда рассчитывают **индекс корреляции** .

**F-тест** используетсядля проверки качества уравнения регрессии, т.е. статистической значимости самого уравнения и коэффициента детерминации.

 – модель статистически НЕ значима,

 – модель статистически значима.

**F-критерий Фишера (F-статистика)**:



при выполнении условий Гаусса-Маркова 1-5 имеет распределение Фишера , где *m* – число параметров модели,  – уровень значимости (вероятность отвергнуть модель, когда она статистически значима).

Если , то нулевую гипотезу  следует отклонить, т.е. принять модель и  статистически значимыми и надежными.

### Критерии выбора модели

Основным недостатком  является то, что при усложнении модели он возрастает и поэтому не может служить достоверным критерием выбора одной модели из нескольких.

**Скорректированный коэффициент детерминации** учитывает число степеней свободы модели:

,

но не учитывает линейность/нелинейность модели.

Также существует несколько *информационных критериев* выбора модели с учетом числа параметров.

**Информационный критерий Акаике** (Akaike information criterion):

,

где *m* – количество параметров в модели.

Абсолютное значение  не имеет смысла, данный критерий может служить лишь для сравнения моделей. Чем меньше величина , тем лучше модель соответствует исходным данным.

**Критерий Шварца** (Schwarz Criterion), или Байесовский информационный критерий (Bayesian information criterion):

.

Чем меньше величина , тем модель лучше, причем критерий Шварца более «чутко» относится к увеличению числа параметров модели, чем Акаике.

Оба критерия  и  рекомендуется использовать для задач большой размерности ().

Для задач малой размерности был предложен **скорректированный критерий Акаике**:

.

Какой из перечисленных критериев использовать, и как учитывать нелинейность модели, остается на усмотрение эксперта.

### Проверка качества оценок параметров

При выполнении условий Гаусса-Маркова можно проверить гипотезу о статистической значимости оценок параметров, т.е. их отличи от нуля:

 – параметр статистически НЕ значим,

 – параметр статистически значим.

Для проверки этой гипотезы строится двухсторонняя **t-статистика**:

,

где  – стандартные ошибки параметров.

Для парной линейной регрессии:

;

.

Для нелинейных моделей вместо *X* необходимо подставлять соответствующие факторы (1/*X*, ln*X*).

Значение t-статистики сравнивается с критическим значением распределения Стьюдента . Если , то параметр считается статистически значимым.

Используя t-статистику можно также построить **доверительный интервал** значений параметра :



и доверительный интервал прогнозного значения регрессии при заданном *X* = *x* (M[*Y*|*X* = *x*]):

.

## Пример выполнения работы

Исследуется зависимость распространенности сети Интернет от общего уровня потребления населения по России. Требуется определить ожидаемый уровень доступа в Интернет по Самарской области при увеличении потребления на 2%.

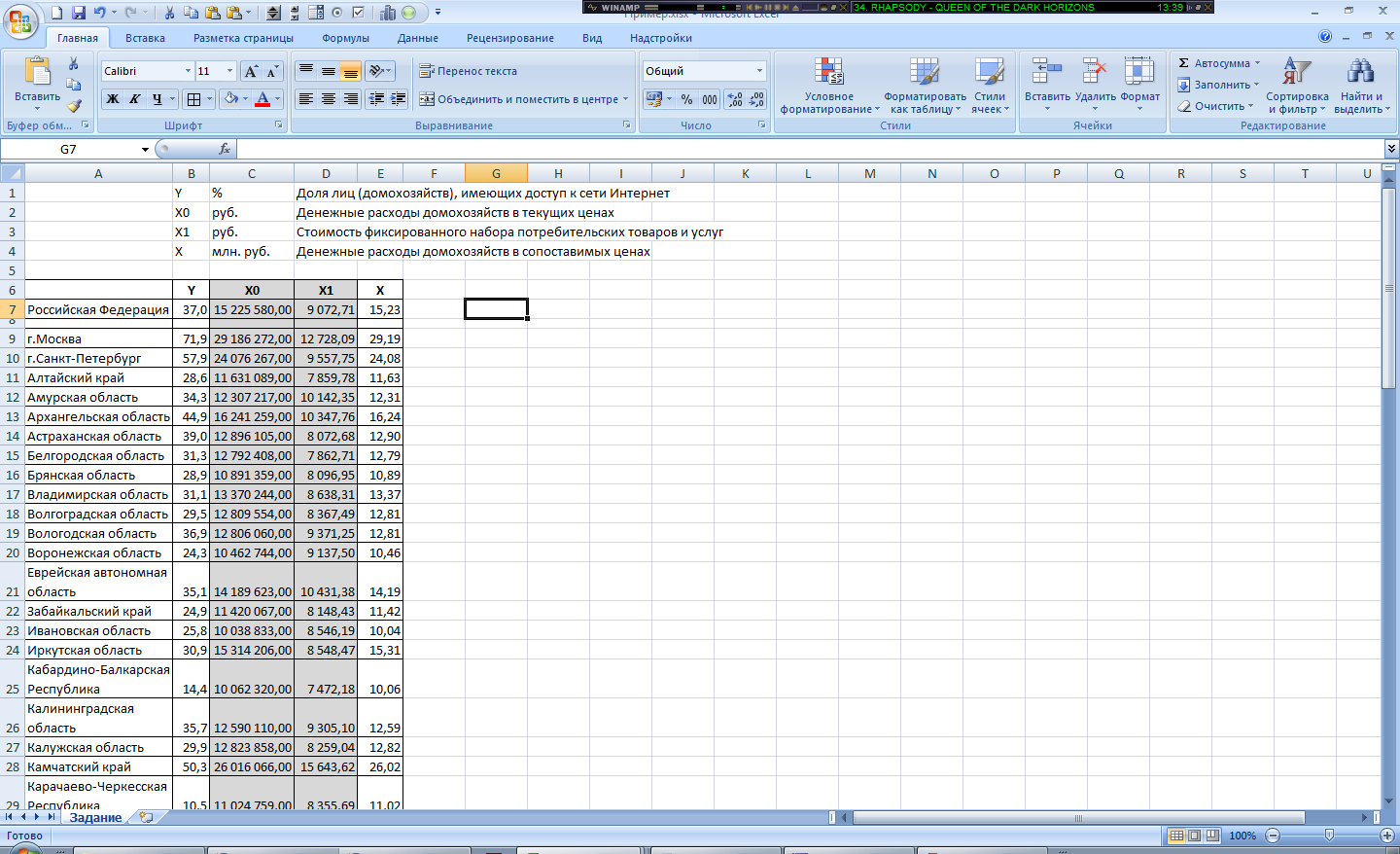
Для исследования были выбраны следующие статистические показатели[[1]](#footnote-1):

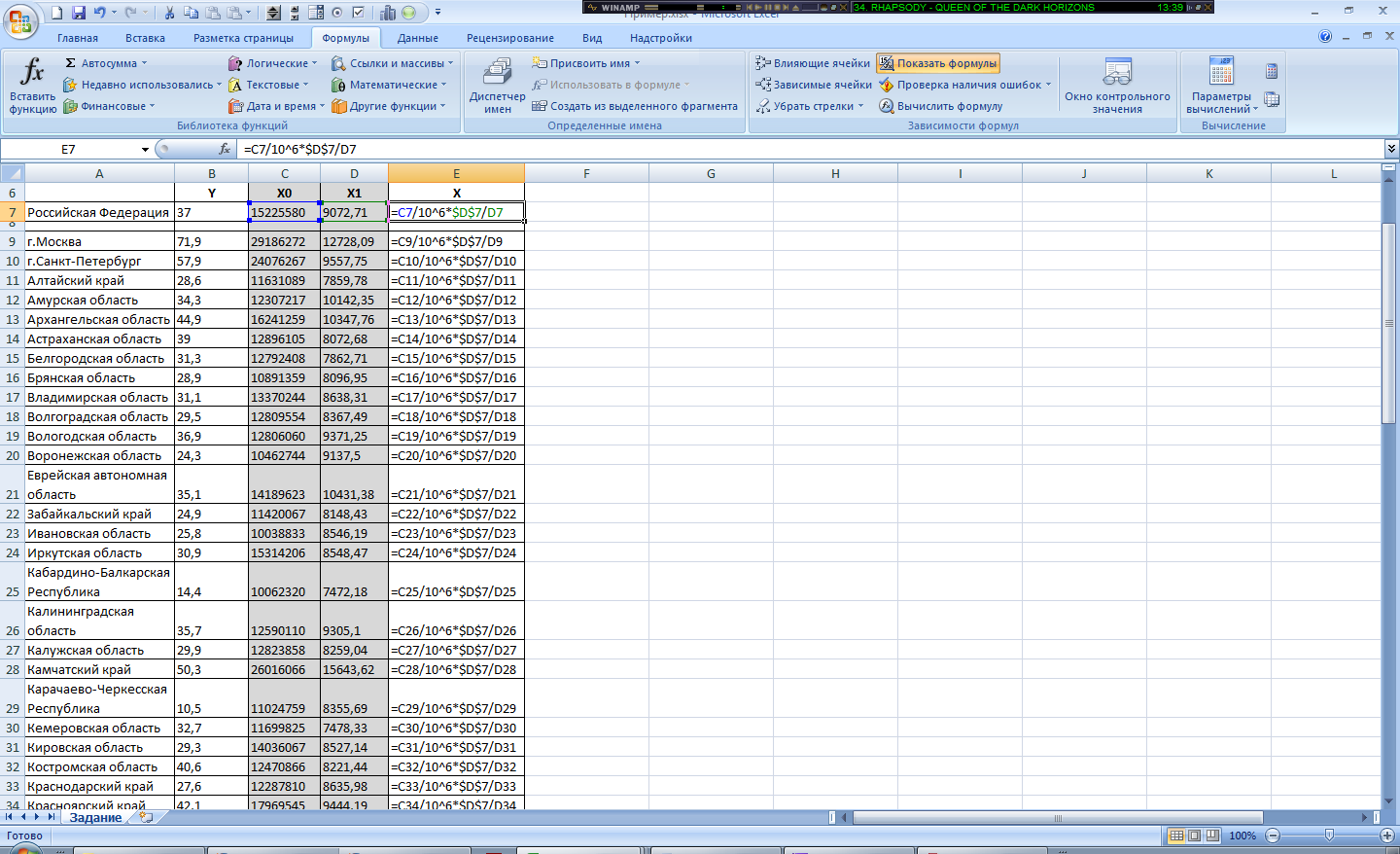
* доля домохозяйств, имеющих доступ к сети Интернет, на конец года, %;
* совокупные денежные расходы домохозяйств за год, в текущих ценах, руб.

Значения показателей за 2011г. доступны для большинства субъектов Российской Федерации ().

Однако значения совокупных денежных расходов не являются сопоставимыми для различных регионов из-за разного уровня цен. Чтобы перевести показатель из текущих цен в сопоставимые, воспользуемся дополнительными сведениями о текущей стоимости фиксированного набора потребительских товаров и услуг, руб.

Таким образом, хотя модель строится для двух показателей, исходные статистические данные содержат три выборки: X0 и X1, на основе которых рассчитывается X, и Y.





 – зависимая (эндогенная) переменная, регрессор;

 – независимая (экзогенная) переменная, фактор.

В качестве эталонного уровня цен выбрано значение  в среднем по России. В других расчетах данные по РФ не используются, поскольку являются усреднением всех остальных значений.

### Визуальный анализ исходных данных.

Построим график зависимости *Y* от *X*.

|  |  |
| --- | --- |
|  | Используйте диаграмму типа «Точечная», без линий. Проверьте правильность выбора осей, выберите удобный масштаб. Включите вертикальные линии сетки, отметьте названия и единицы измерения осей. |

График позволяет предположить наличие положительной линейной взаимосвязи между исследуемыми показателями, однако разброс значений вокруг предполагаемой линии регрессии достаточно велик. Выдвинуть какие-либо предположения о характере нелинейной зависимости в данном случае по графику сложно: отсутствуют видимые минимумы или максимумы, нельзя сказать выпуклая или вогнутая должна быть функция.

### Корреляционный анализ.

Чтобы проверить возможные виды зависимостей, рассчитаем парные коэффициенты корреляции между *Y* и *X*, *Y* и *X*2, *Y* и 1/*X*, Y и ln*X*, ln*Y* и *X*, ln*Y* и ln*X*. Каждый из них соответствует определенной двухпараметрической модели:

|  |  |
| --- | --- |
| Y|X |  |
| Y|X2 |  |
| Y|1/X |  |
| Y|lnX |  |
| lnY|X |  |
| lnY|lnX |  |

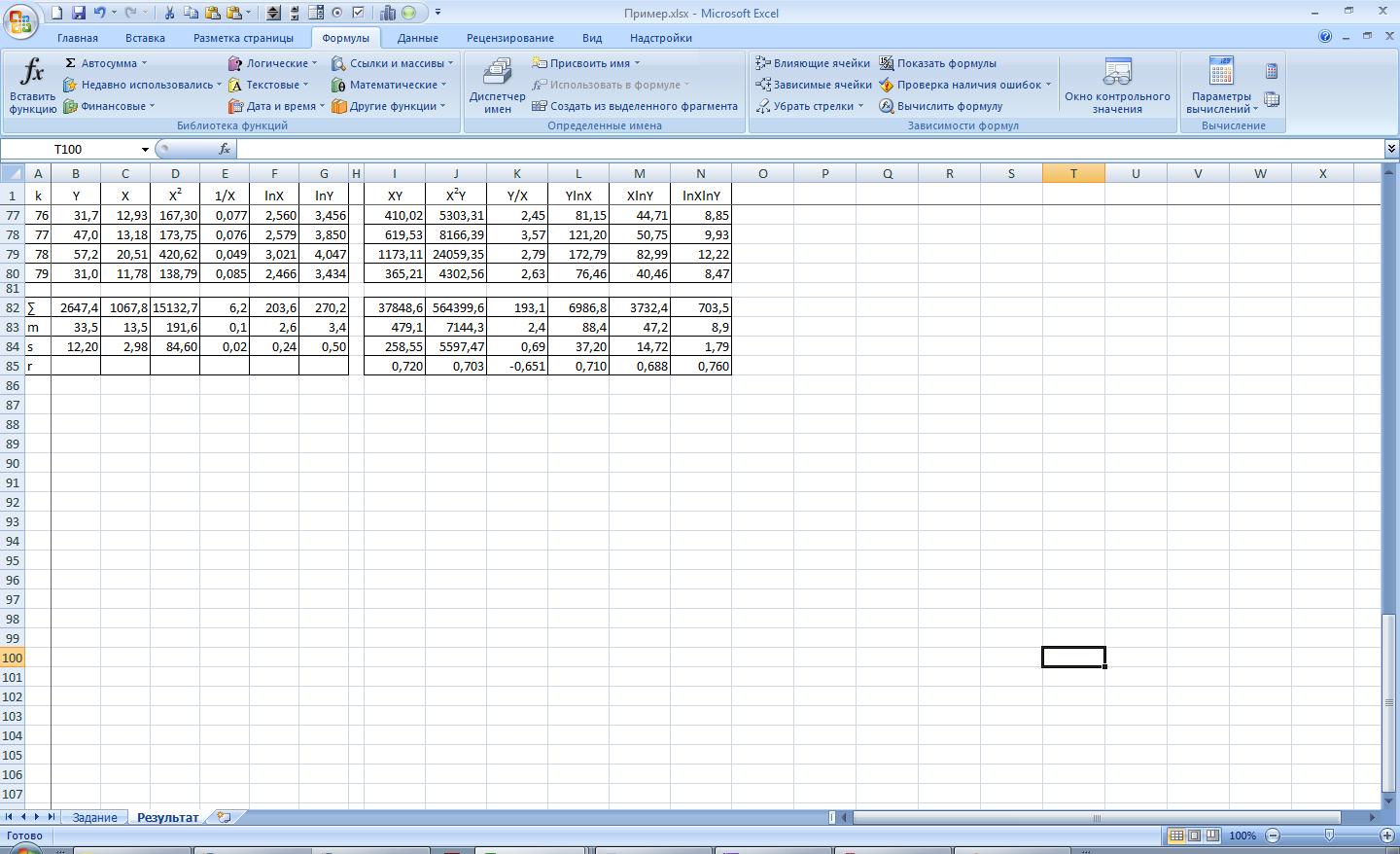
Обратите внимание, что Y|X2 соответствует квадратичной, а не полной параболической зависимости .

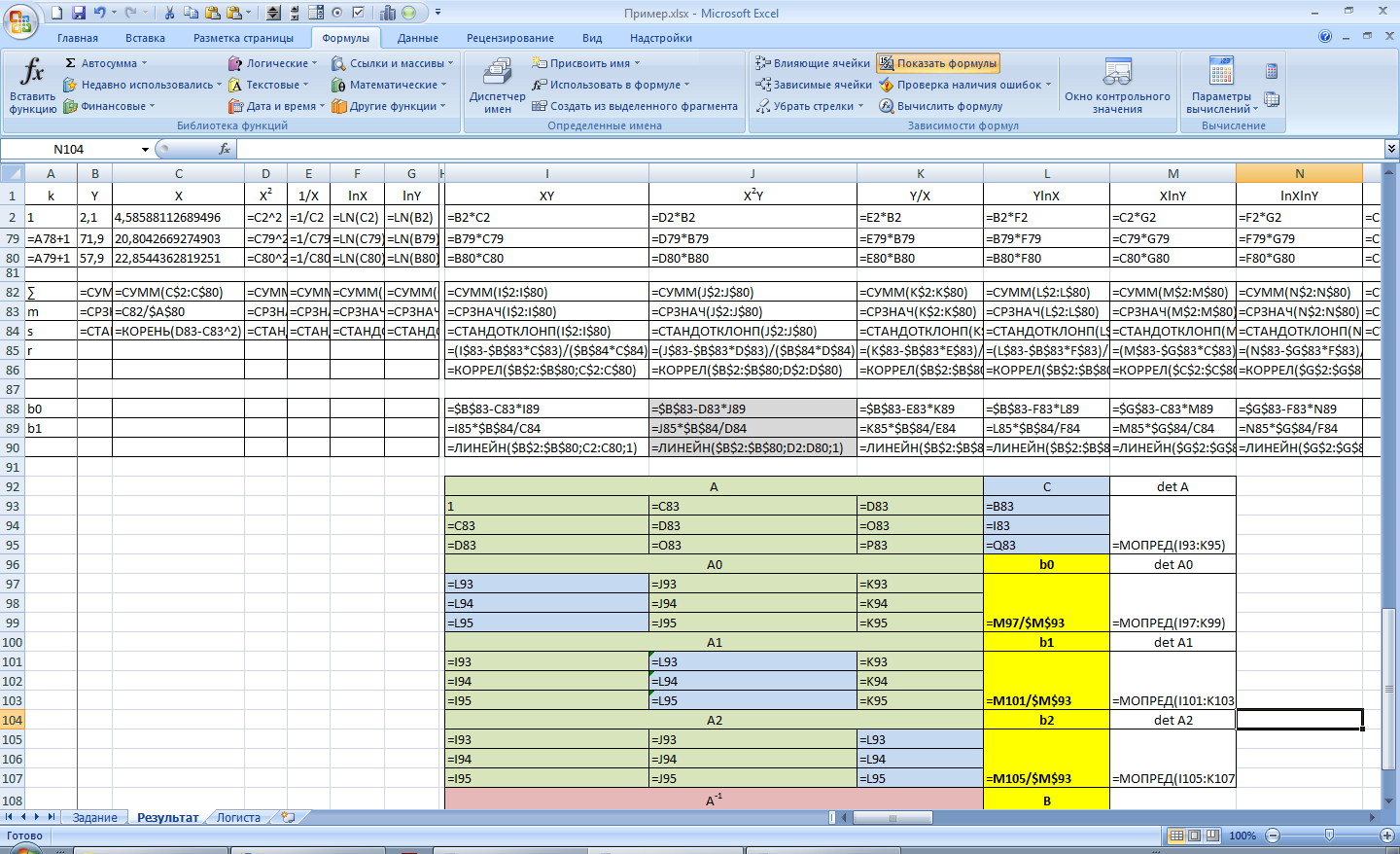
Воспользуемся следующей формулой коэффициента корреляции:

, 

где  – смещенные оценки среднеквадратического отклонения.

|  |  |  |
| --- | --- | --- |
|  |  | =СРЗНАЧ(диапазон X) |
|  | =СТАНДОТКЛОН.В(диапазон X) |
|  | =КОРРЕЛ(диапазон X; диапазон Y) |





Таким образом, для всех предполагаемых зависимостей коэффициент корреляции по модулю близок к 0,7 (заметная линейная связь). Наибольшее значение коэффициента корреляции у показательной зависимости (0,760), наименьшее – у обратной (гиперболической) (0,651), но отличия невелики. Следовательно, все эти зависимости потенциально могут дать неплохую модель.

### Идентификация моделей.

Рассмотрим возможные варианты вида регрессии:

1. линейная ;
2. параболическая ;
3. обратная (гиперболическая) ;
4. логарифмическая ;
5. показательная (экспоненциальная) ;
6. степенная .

Выполним идентификацию предложенных моделей.

Параболическая, гиперболическая и логарифмическая модели линейны по параметрам и могут быть идентифицированы с помощью МНК.

Чтобы идентифицировать нелинейные по параметрам модели выполним их линеаризацию:

1. ;
2. .

Рассчитаем оценки параметров моделей  по формулам коэффициентов регрессии.

Для регрессий с двумя параметрами:

.

|  |  |  |
| --- | --- | --- |
|  | Проверьте себя! | |
| линейная регрессия | =ЛИНЕЙН(диапазон Y; диапазон X) |
| показательная регрессия | =ЛГРФПРИБЛ(диапазон Y; диапазон X) |
| Эти функции можно использовать в одной ячейке, тогда они возвращают значение b1, либо как формулы массива, чтобы получить оба параметра. Для этого выделите формулу с ячейкой и соседнюю справа от нее, нажмите F2, а затем Ctrl+Shift+Enter. В ячейке с формулой отобразится b1, а в соседней – b0.  Другие модели (кроме гиперболической) можно получить, добавив на график исходных данных линии тренда. | |

Через  рассчитаем оценки  параметров  для показательной и степенной регрессий:

; .

|  |  |
| --- | --- |
|  | Экспонента x вычисляется с помощью функции =EXP(x) |

Для параболической регрессии с тремя параметрами:

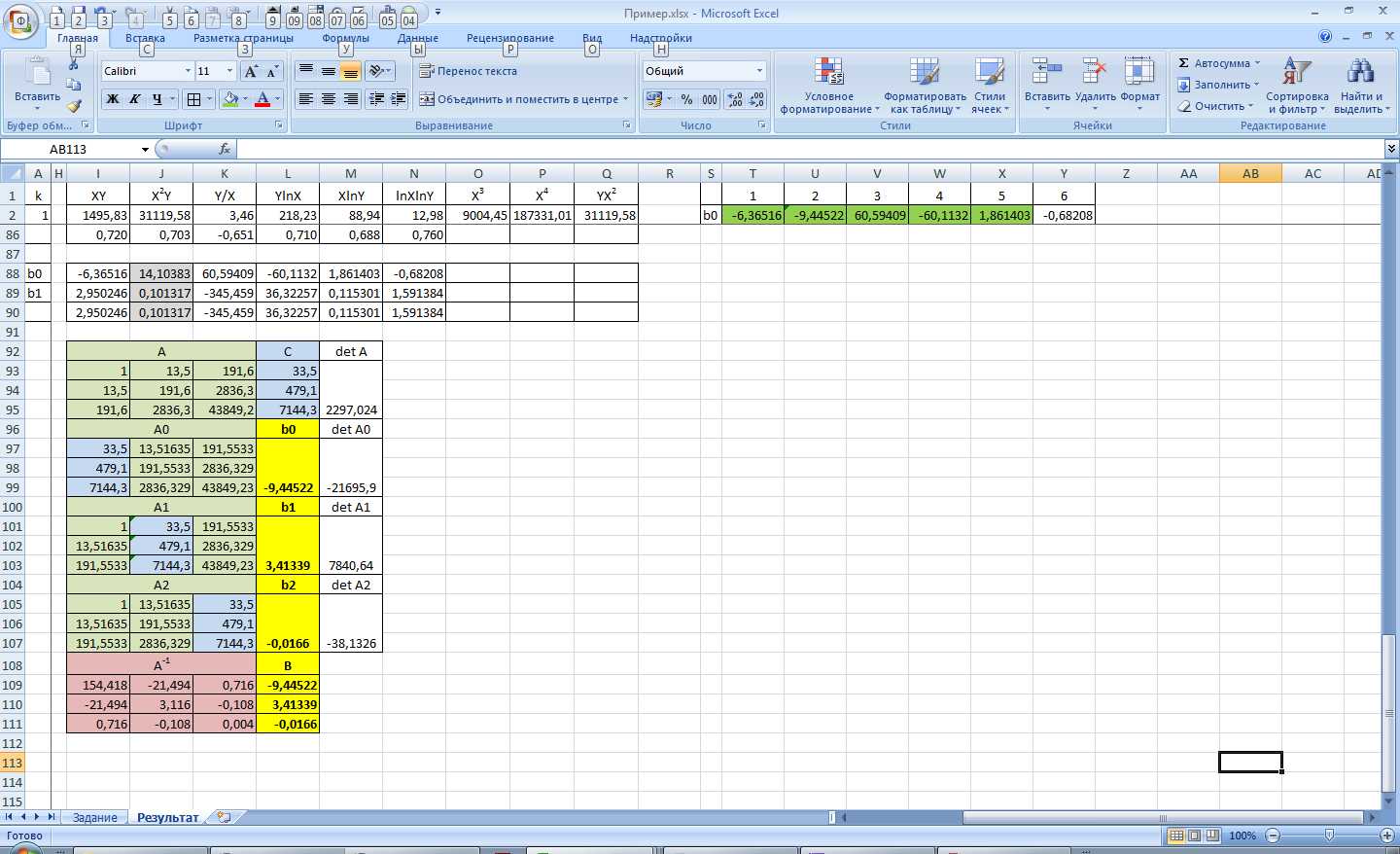


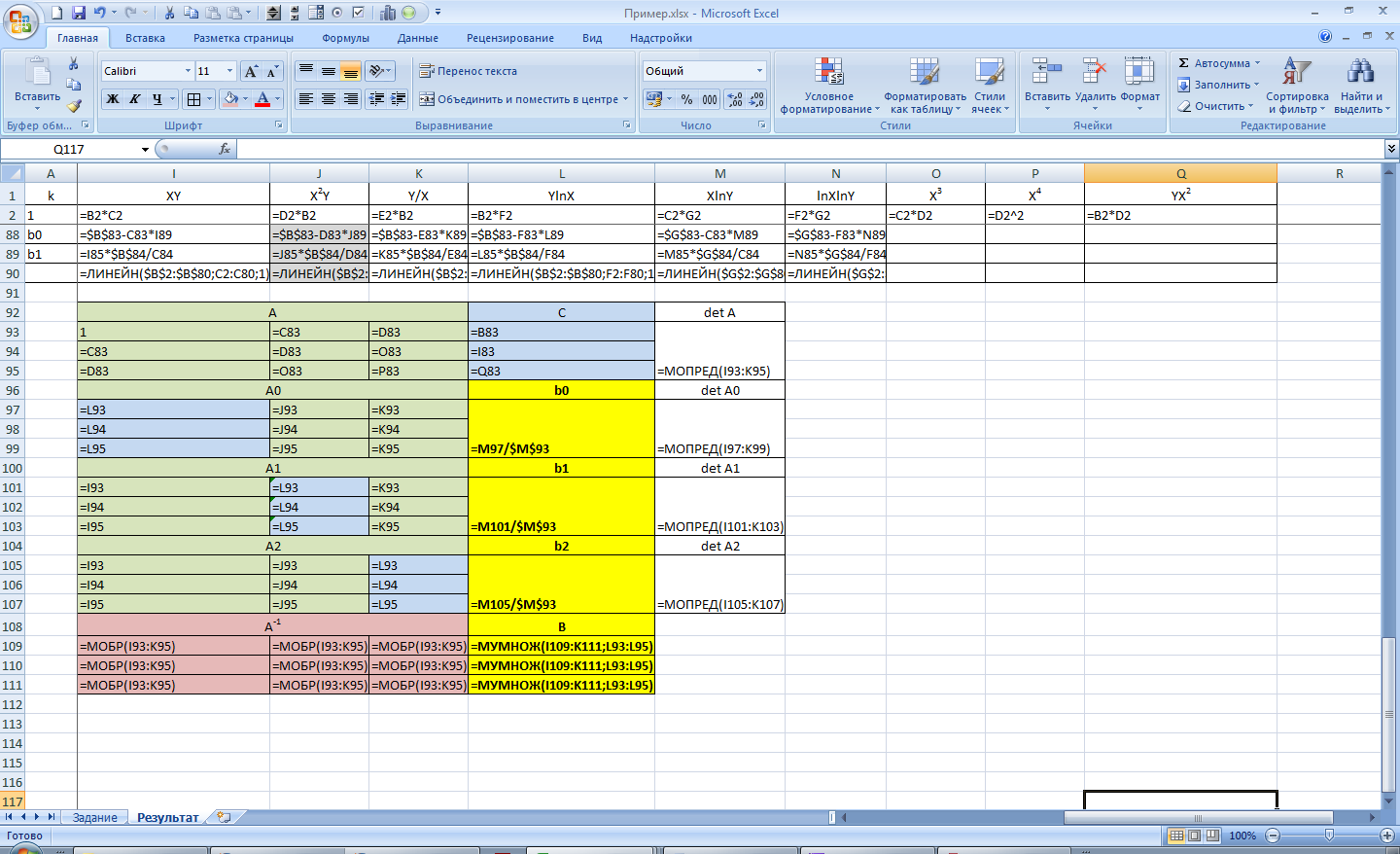
Или, по методу определителей:

,

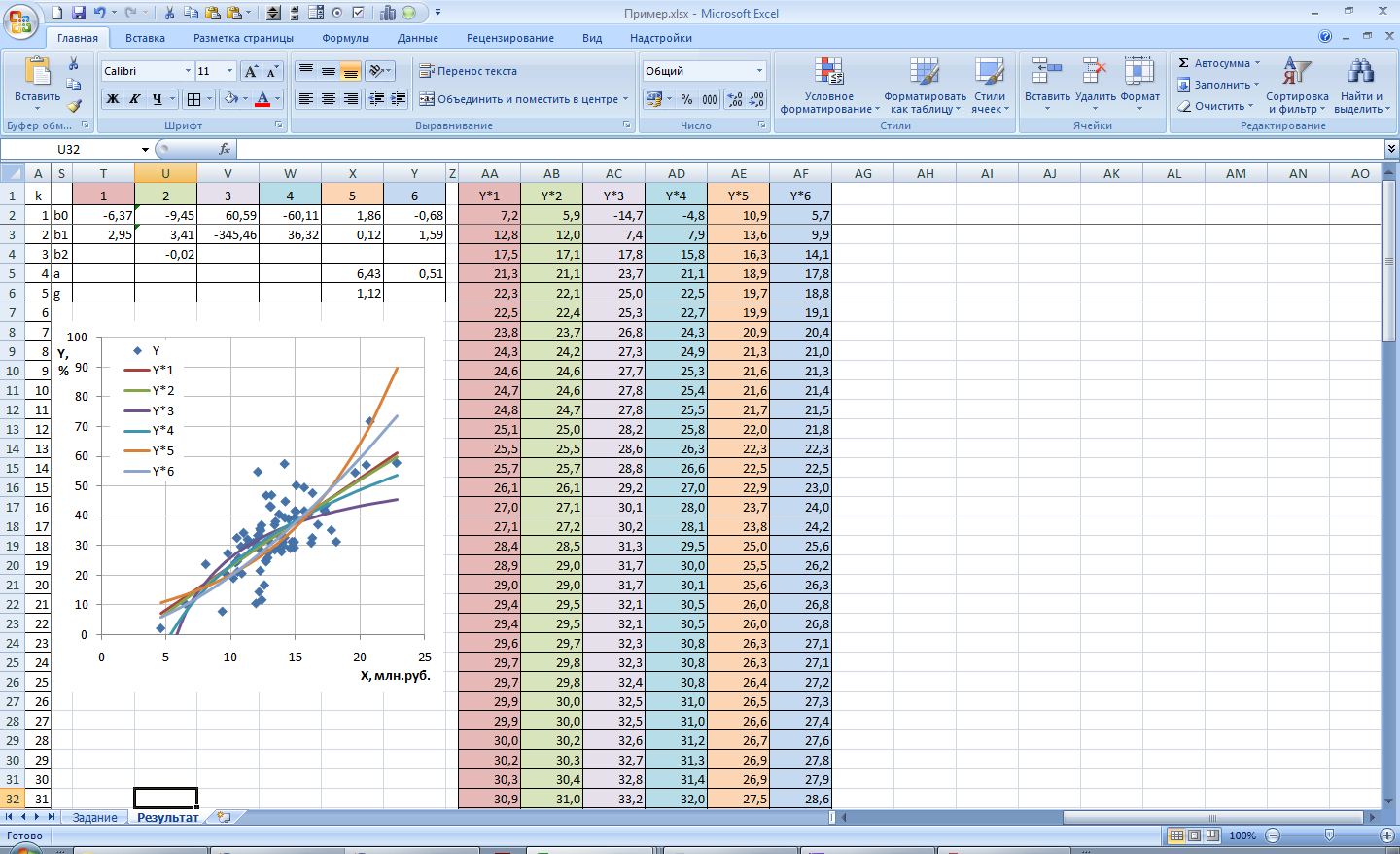
где  – определитель матрицы *А*, , ,  – частные определители (в матрице *A* соответствующий столбец заменен на *C*).

|  |  |
| --- | --- |
|  | Обратную матрицу можно рассчитать с помощью функции МОБР(матрица). Результат этой функции является массивом (выделить область результата, начиная с ячейки с формулой, нажать F2, нажать Ctrl+Shift+Enter).  Умножение векторов и матриц осуществляется функцией МУМНОЖ(вектор, матрица). Ее результатом также является массив. |
| Для расчета определителя матрицы используйте функцию МОПРЕД(матрица). |





Рассчитаем модельные уровни регрессии и покажем их на графике:



|  |  |
| --- | --- |
|  | Чтобы построить гладкие графики моделей, отсортируйте исходные данные по X. Выделите массивы для X и Y (с заголовками), на вкладке «Главная»→ «Сортировка и фильтр» → «Настраиваемая сортировка». В окне выберите сортировку по столбцу X. Остальные данные должны рассчитаться автоматически.  В данном случае порядок появления пар X и Y не важен (все наблюдения являются равнозначными), хотя в некоторых случаях он имеет значение (наблюдения не равнозначны). |

Таким образом, получены следующие модели регрессии:

1. ;
2. ;
3. ;
4. ;
5. ;
6. .

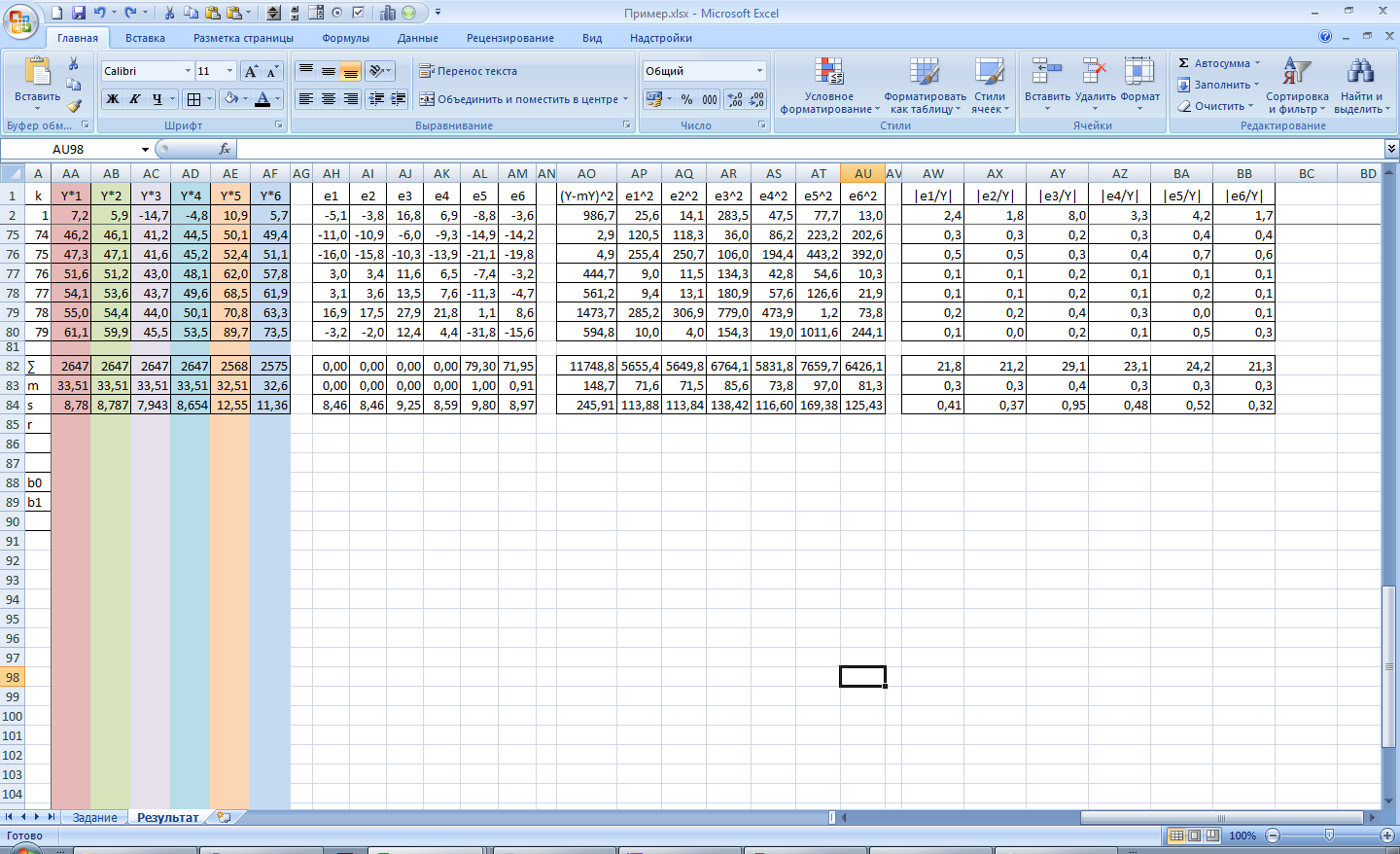
### Оценка общего качества моделей.

Для оценки качества модели необходимо вычислить случайные остатки:



для каждой модели (в том числе и для мультипликативных).

Если вычисления выполнены правильно, то для моделей с аддитивной помехой . Для мультипликативных моделей  из-за асимметричности логнормального закона распределения.



Вычислим коэффициент детерминации по формуле:

.

|  |  |
| --- | --- |
|  | Значение R^2, которое можно отобразить на диаграмме для линии тренда, вычисляется по другой формуле, как квадрат коэффициента корреляции *r*2. Для моделей, линейных по параметрам, эти величины совпадут, а для нелинейных *R*2 < *r*2. |

Проверим его статистическую значимость с помощью F-критерия:

,

где *m* – число параметров в модели, в нашем случае 2 или 3.

Чем больше значений в выборке и чем выше , тем больше значение F-критерия и тем надежнее модель.

F-критерий должен превышать с критическое значение F-распределения Фишера  с уровнем значимости  и степенями свободы  и :

.

Такую проверку можно осуществлять только при выполнении всех условий Гаусса-Маркова 1-5 и для моделей, линейных по параметрам. Также можно проверить статистическую значимость линеаризованных форм нелинейных моделей, используя вместо *R*2 квадрат коэффициента корреляции, рассчитанного ранее, *r*2.

 обычно выбирают равным 0,1; 0,05; 0,01; 0;001, учитывая объемы выборок и уровень зашумленности модели. В данном случае объем выборки приближается к 100, поэтому, хотя шум достаточно велик, назначим .

Имея в своем распоряжении компьютер, можно вычислить  из условия , т.е. найти вероятность, с которой  не является статистически значимым.

|  |  |  |
| --- | --- | --- |
|  | *Fкр*. | =F.РАСП.ОБР(1–;;) |
|  |  | =F.РАСП (*F*;;) |
|  |  | |

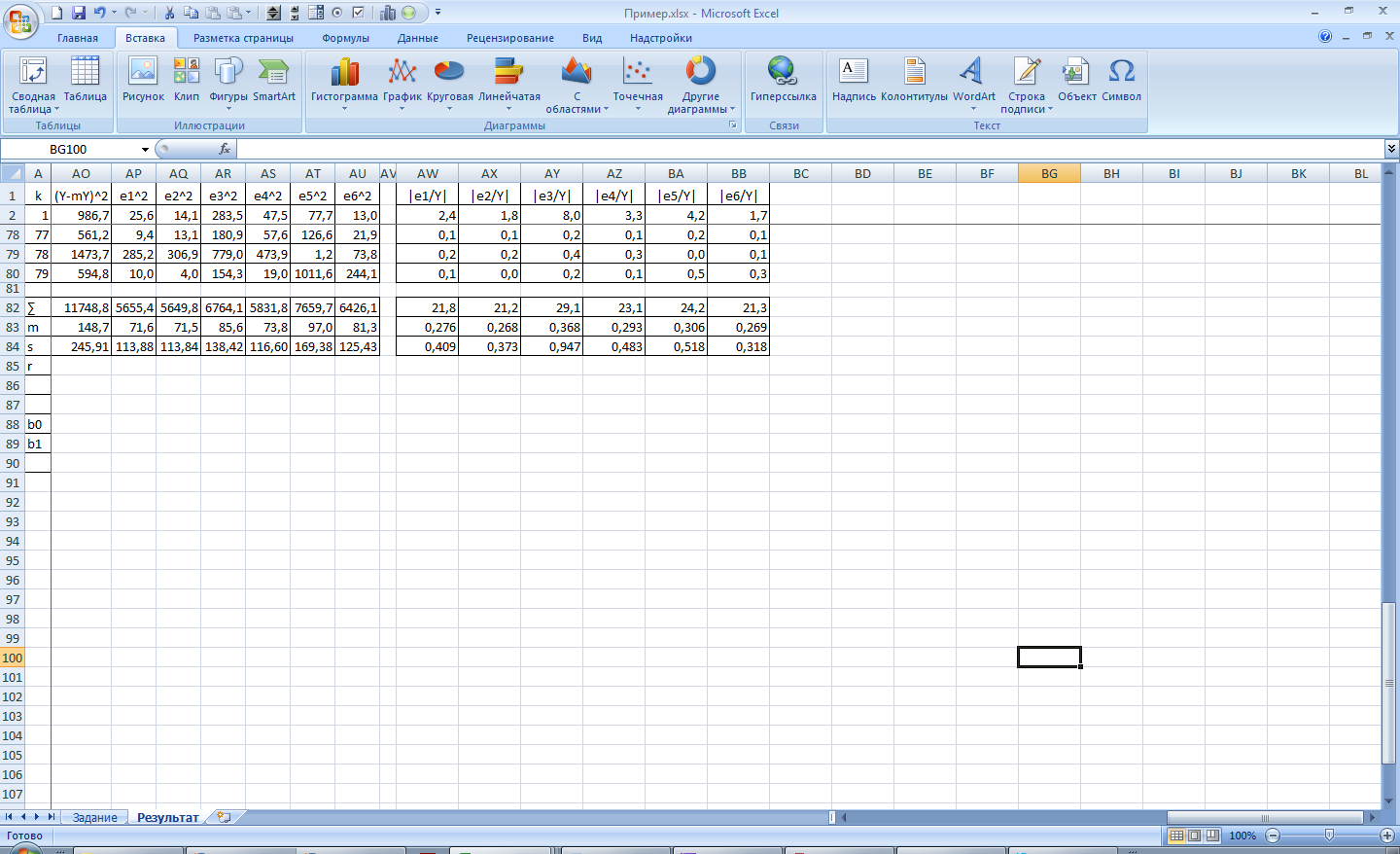
Кроме того, вычислим MAPE-оценку, которая показывает среднее процентное отклонение реальных данных от модельных:

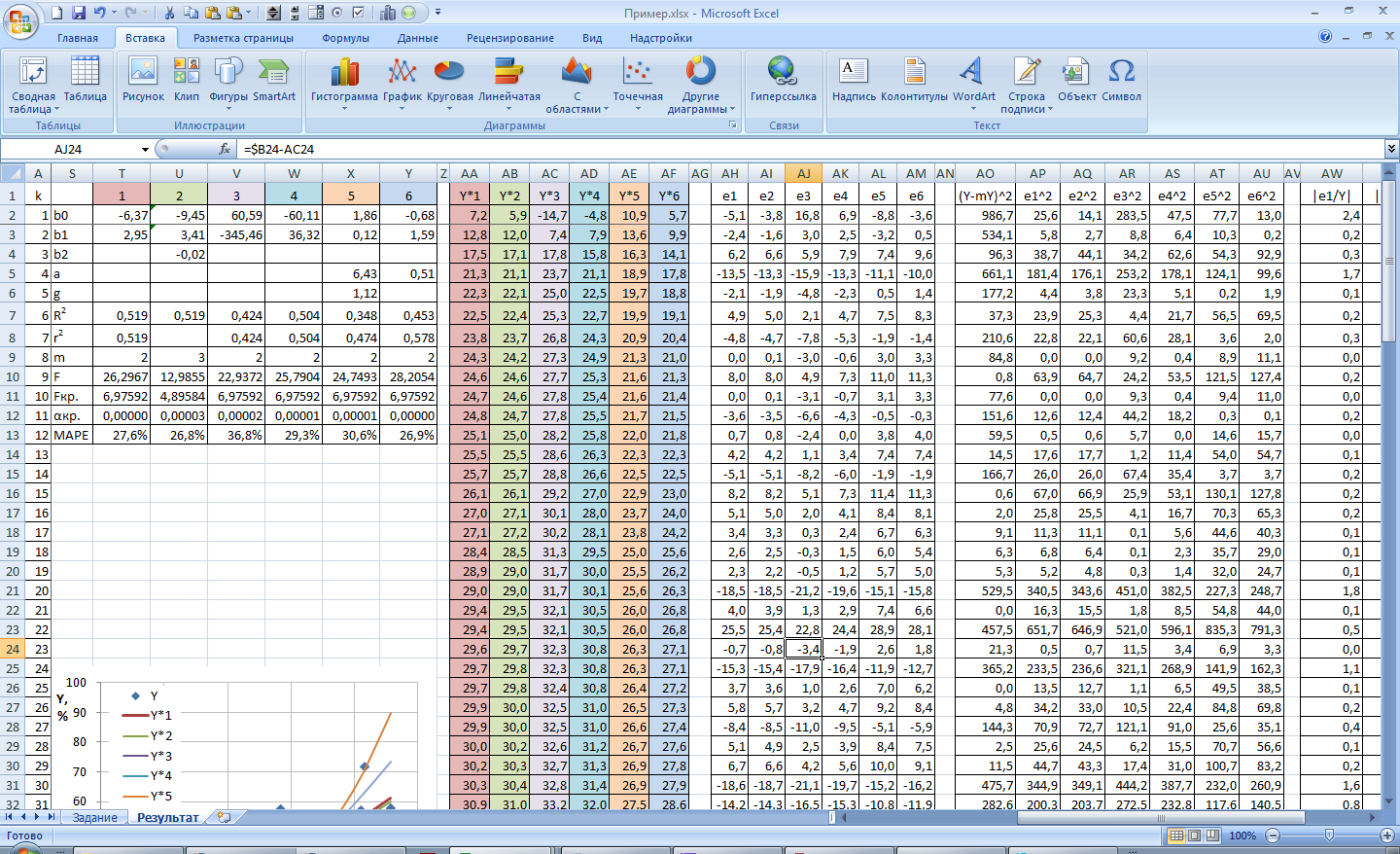
.

Чем меньше MAPE-оценка, тем точнее модель. Допустимой будем считать MAPE < 30%, удовлетворительной MAPE < 20%, хорошей MAPE < 10%.

|  |  |
| --- | --- |
|  | Модуль числа x вычисляется с помощью функции =ABS(x). |

MAPE-оценка не имеет статистического смысла, но зачастую бывает более наглядной. Кроме того, стоит обратить внимание на MAPE в мультипликативных моделях, т.к. в них ошибка задается как доля от значений показателя.





По примерному правилу *R*2>0,5 и с использованием общей формулы для расчета *R*2 получаем, что только линейная (1) и логарифмическая (4) модели обладают приемлемой точностью. Однако в п.2 для всех линеаризованных моделей коэффициент корреляции был в районе 0,7 (для линейных по параметрам моделей *R*2 равен квадрату *r*, а для нелинейных – нет).

В то же время, F-критерий показывает, что в данном случае все модели можно признать статистически значимыми с вероятностью ошибки (первого рода) менее 0,001. Причиной тому служит большой объем исходной выборки.

MAPE-оценка для всех моделей находится в районе 25-35%, т.е. достаточно велика. Отметим, насколько различны соотношения MAPE и *R*2: наименьшая MAPE у 2 и 6 моделей, а наибольший *R*2 у 1 и 4. Какой из критериев считать более важным, зависит от целей исследования и характера исходных данных (что важнее – абсолютная ошибка или относительная, наблюдается ли гетероскедастичность помехи и др.) .

В данном случае можно однозначно отметить низкое качество гиперболической модели (3) и экспоненциальной модели (5).

### Выбор наилучшей модели.

Для выбора наилучшей модели рассчитаем скорректированный коэффициент детерминации, информационный критерий Ака́ике и критерий Шварца:

,

,

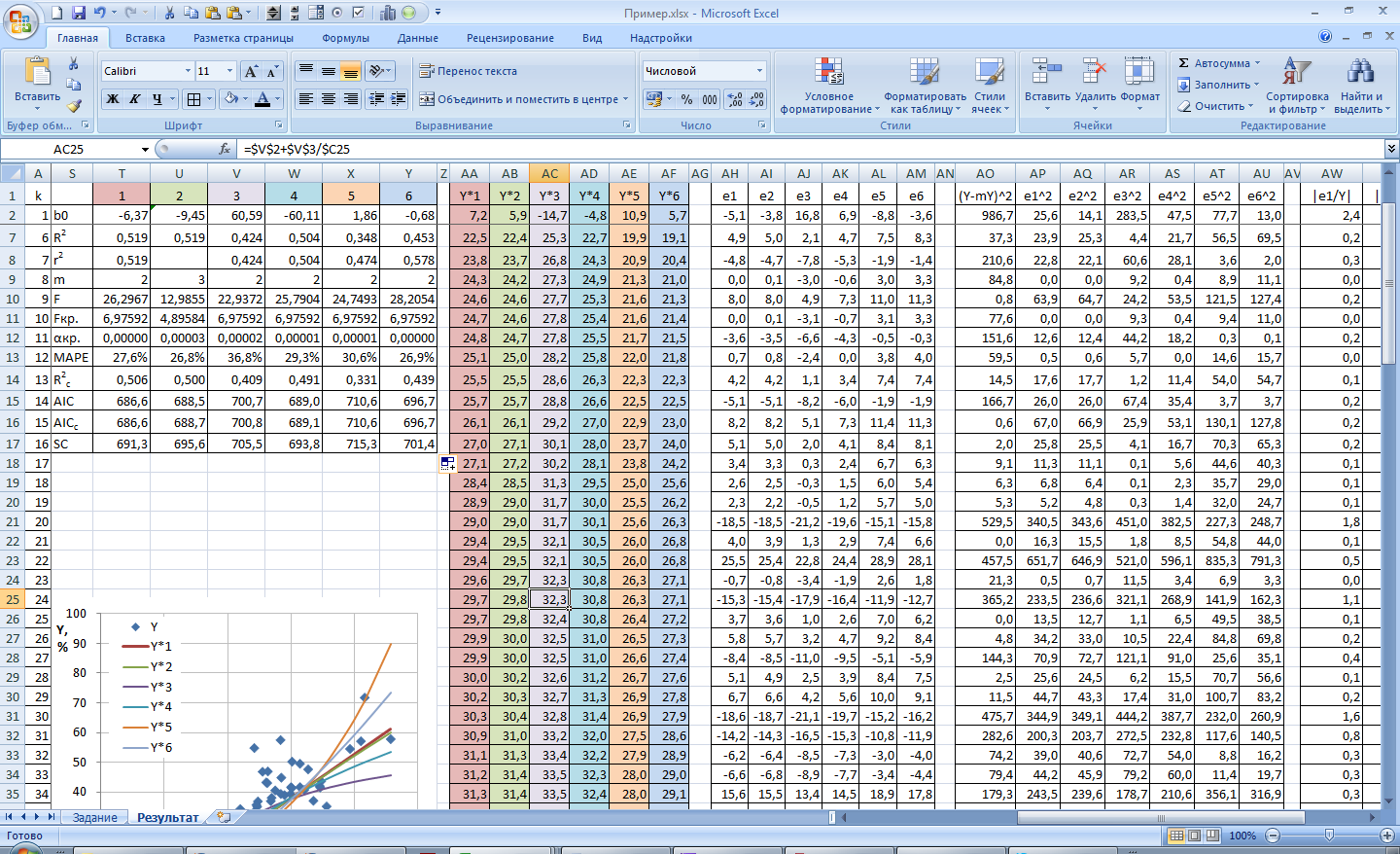
.

Если число наблюдений мало (*n*/*m* < 40), критерий Шварца использовать не рекомендуется, а критерий Акаике необходимо скорректировать:

.

В данном случае это соотношение нарушается для параболы , но с небольшой натяжкой можно использовать основные критерии и для нее. Для наглядности, в данном примере рассчитаем все критерии.

|  |  |
| --- | --- |
|  | В этих формулах часто приходится ссылаться на n – число наблюдений. Чтобы не вводить его вручную, но и не приходилось искать нужную ячейку на листе, можно задать для нее имя. Выделите ячейку, содержащую n, и в окне «Имя» (рядом со строкой формулы) впишите “n”. Теперь вместо абсолютной ссылки на ячейку можно использовать назначенное имя. |



Скорректированный *R*2 всегда меньше обычного. Разница между скорректированным и обычным критерием Акаике в данном случае очень мала. Критерий Шварца, который всегда (для n > 7) превышает критерий Акаике и более чувствительно относится к включению дополнительных параметров в модель, в данном случае не дает ощутимо отличающихся результатов для параболической модели (2), которая содержит 3 параметра.

Формально, все критерии указывают на наилучшее качество линейной модели (1), а наихудшим качеством обладают гиперболическая модель (3) и экспоненциальная модель (5).

Формальность такого выбора опять же связана с тем, что две модели из шести являются нелинейными по параметрам, и сумма квадратов отклонений USS для них будет завышенной. Из 4 линейных по параметрам моделей наилучшим качеством однозначно обладает линейная модель (1).

Таким образом выбор осуществляется между линейной (1) и степенной моделью (6). Несмотря на неудовлетворительные значения некоторых критериев для последней, соответствующая ей линеаризованная модель является наилучшей (*r*2= 0,58). Чтобы сделать правильный выбор, обратимся к сути рассматриваемых параметров.

Регрессор *Y* представляет собой долю населения (домохозяйств), имеющих доступ к сети Интернет в различных субъектах РФ. Очевидно, что он изменяется в пределах от 0 до 100%. Фактически, на данный момент он варьируется от 2,1% (Республика Ингушетия) до 71,9% (Москва) и, скорее всего, в настоящее время 100%-ный охват населения недостижим по техническим причинам, т.е. показатель ограничен сверху уровнем порядка 90-95%.

Фактор *X –* общий объем расходов домохозяйств, для которого устранена разница в ценах между субъектами РФ, характеризует совокупный спрос домохозяйств. Данный показатель не может быть отрицательным, но установить его верхнюю границу, пожалуй, невозможно.

Очевидно, что при увеличении совокупного спроса все большее число домохозяйств получает доступ к сети Интернет. Но при приближении к 100%-му уровню темпы роста будут замедляться, поскольку остается все меньше людей, которые нуждаются в Интернете и имеют техническую возможность доступа. При очень низком объеме расходов (ниже определенного уровня, например прожиточного минимума) доля пользователей сети будет близка к 0 и практически не будет изменяться (небольшой процент, скорее всего, все же останется – те, кому Интернет необходим для работы, учебы).

Таким образом, имеем картину: почти постоянный уровень Y при низких значениях X, рост Y при росте X, замедление роста при приближении Y к максимуму. Такая динамика называется **логистической**. Один из вариантов (не единственный!) логистической кривой (логисты, S-образной кривой), в сравнении с моделями (1) и (6) показан на графике.

Построение логистических кривых выходит за рамки данной лабораторной работы, поскольку они являются существенно нелинейными по параметрам. Однако можно заметить, что середина логисты близка к прямой линии, начало (нижняя половина) представляет собой выпуклую вниз кривую (ветвь параболы, экспонента, степенная функция), а верхняя половина – вогнутую кривую (другая ветвь параболы, гипербола, сумма константы и экспоненты с отрицательным показателем).

Следовательно, модель (6) в большей степени соответствует малым значениям X и Y, а модель (1) – близким к средним значениям. При больших значениях X (более 20 млн. руб.) более грубой является степенная функция, поскольку на слишком быстро возрастает. Поэтому решение об адекватности модели зависит от того, какой диапазон X нас интересует, т.е. от цели моделирования.

В задании сказано, что требуется определить ожидаемый уровень доступа в Интернет по Самарской области при увеличении общего объема расходов населения на 2%, т.е. для уровня расходов населения, равного 17,6 млн.руб., что существенно превышает средний уровень (на 30% *mX*, или 1,4*sX*), поэтому предпочтение стоит отдать линейной модели. Но если бы нас интересовала, к примеру, Республика Мордовия (X = 10,2 млн.руб.), следовало бы выбрать степенную зависимость.

Также можно отметить, что степенная функция проходит через начало координат, т.е. при нулевых расходах ожидаемый уровень доступа к сети Интернет равен нулю, но при любых малых расходах Y будет положительным. Прямая линия, напротив, пересекает ось X в точке 2,15 млн. руб., т.е. при расходах менее этой суммы ожидаемый уровень доступа к сети Интернет будет отрицательным. Также по линейной модели 100% уровень доступа будет достигнут при X = 36,1 млн.руб., а по степенной – 27,7 млн.руб.

### Оценка качества идентификации параметров.

Выполним оценку качества и построение доверительных интервалов оценок параметров линейной модели (1), выбранной в качестве наилучшей в п.5.

Заметим, что в реальности эта проверка должна осуществляться для каждой построенной модели вместе с общей оценкой качества модели, но в лабораторной работе она была вынесена в отдельный пункт для сокращения объема расчетов. Для нелинейных моделей все расчеты проводятся для линеаризованной формы модели, а доверительный интервал исходных параметров можно рассчитать по связывающим эти параметры формулам.

В первую очередь необходимо рассчитать стандартные ошибки параметров модели:

; .

Затем для каждого параметра рассчитывается t-статистика:

.

Для нахождения критического уровня  зададимся уровнем значимости . Если

,

то соответствующий параметр признается статистически значимым.

Как и для F-критерия, можно рассчитать  для полученных t-статистик (вероятность принять параметр статистически значимым, тогда как он таковым не является).

|  |  |  |
| --- | --- | --- |
|  | *tкр*. | =СТЬЮДЕНТ.ОБР(1–/2;) |
|  |  | =СТЬЮДЕНТ.РАСП (abs();) |
|  |  | |

Наконец, определим полуразмах доверительного интервала для каждого параметра:

.

Можно показать на графике, насколько изменится модель, если параметры будут равны границам своих интервалов.



В данном случае с доверительной вероятностью  параметр *b*1 является статистически значимым, а параметр *b*0 – нет. Вероятность ошибочно принять параметр *b*0 статистически значимым составляет 0,160 – достаточно большое значение, а для *b*1 эта вероятность равна к нулю с точностью до пятого знака. Это означает, что наклон прямой определен достаточно надежно, а ее точка пересечения с осью Y – нет.

Как следствие, доверительный интервал *b*0 накрывает 0. По графику видно, что наклон прямой в пределах доверительного интервала изменяется незначительно, а вот возможный сдвиг по вертикали – очень сильно. Это обстоятельство могло бы служить причиной для отклонения линейной модели.

### Прогнозирование по модели.

Рассчитаем значение фактора *x*, для которого необходимо выполнить прогнозирование.

.

Точечная оценка прогноза рассчитывается простой подстановкой *x* в уравнение регрессии:

.

На данный момент уровень распространенности сети Интернет для Самарской области составляет 41,6, т.е. увеличение объема расходов населения на 2% приведет к среднему увеличению распространенности сети Интернет на 9,8%.

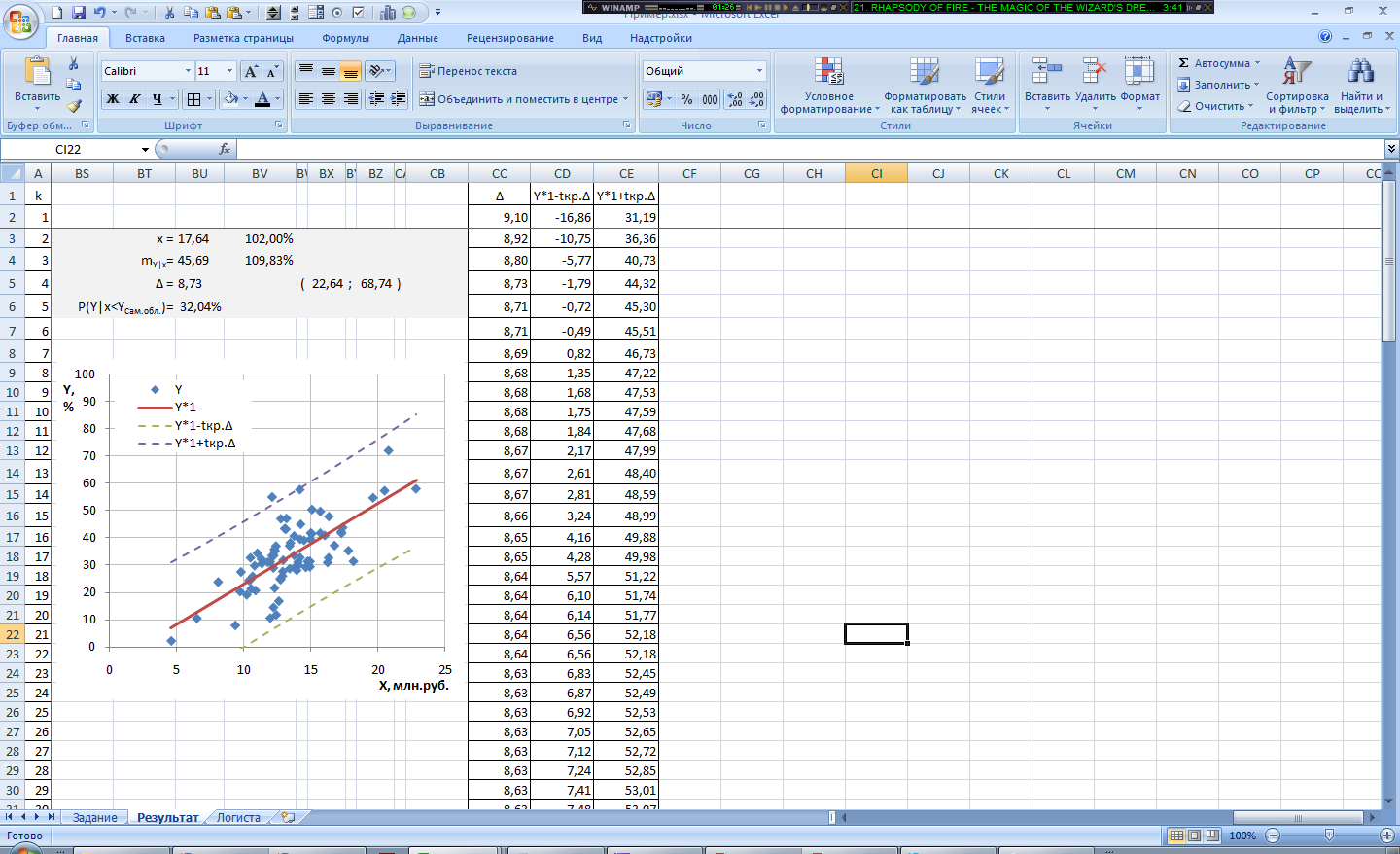
Для построения доверительного интервала прогноза рассчитаем (уровень значимости также равен 0,01):



.

Обратите внимание, что значение  зависит от *x*, для которого рассчитывается интервал, точнее от его удаленности от среднего значения. Поэтому, если построить границы доверительных интервалов для всей прямой, они будут расширяться по краям (чем дальше от среднего *X*, тем менее надежен прогноз).

По той же причине нельзя построить границы доверительного интервала, просто подставив в модель границы доверительных интервалов параметров, как это было показано на графике в п.6.



Таким образом, с вероятностью 99% при увеличении объема расходов населения на 2% уровень распространенности сети Интернет будет составлять от 22,6 до 68,7, т.е. существует довольно большой шанс, что он не изменится или уменьшится. Вероятность этого события составляет 32,0%:

,

где  – значение закона распределения Стьюдента (двухстороннего) с числом степеней свободы .

## Оформление отчета

Отчет по лабораторной работе оформляется в текстовом процессоре MS Word или аналогичном и сдается в печатном виде. Рекомендуется использовать шрифт Times New Roman или Cambria, 14пт., абзацный отступ 1,25см, межстрочный интервал – 1 или 1,5, поля: верхнее 1,5-2см, нижнее 2см, правое 1-1,5см, левое 2,5-3см. Страницы нумеруются снизу по центру, кроме титульного листа.

Отчет должен содержать:

1. Титульный лист.
2. Общее и индивидуальное задание.
3. Основная часть.

Результаты выполнения каждого пункта задания с краткими комментариями и указанием расчетных формул. При необходимости приводить основные промежуточные расчеты. Весь качественный анализ можно перенести в вывод.

1. Вывод.

Привести и сопоставить основные численные результаты работы. Попытаться дать качественное обоснование полученной модели, пояснить смысл ее параметров и функционального вида.

## Контрольные вопросы

1. Каковы основные виды эконометрических моделей и для каких целей они строятся?
2. Перечислите этапы построения эконометрической модели. Насколько строгим является их порядок?
3. В чем отличие теоретического и эмпирического уравнений регрессии?
4. В чем заключается суть МНК?
5. Какие оценки считаются «синими»?
6. Перечислите условия Гаусса-Маркова. Какие из них являются необязательными, а какие следует проверять?
7. Какие виды нелинейных моделей регрессии принято выделять? Приведите примеры.
8. На графике зависимости Y от X наблюдается наличие минимума. Какую модель регрессии можно рекомендовать?
9. Какие проблемы возникают при линеаризации нелинейных по параметрам моделей?
10. В чем отличие адекватности и точности модели? Что обычно подразумевают под проверкой качества уравнения регрессии?
11. Перечислите известные Вам показатели общего качества уравнения регрессии.
12. Охарактеризуйте качество линейной модели с R2=0,554, если исходная выборка содержит 20 значений? 10 значений? 100 значений?
13. Зачем нужны интервальные оценки параметров и прогнозных значений?
14. От чего зависит ширина доверительного интервала параметров?

1. Источники:www.gks.ru, www.fedstat.ru [↑](#footnote-ref-1)