ФГБУН Институт экологии человека СО РАН

**ГБОУ ВПО «Кемеровская государственная медицинская академия»
Министерства здравоохранения и социального развития РФ**

**ФГБУ НИИ Комплексных проблем сердечно-сосудистых заболеваний
СО РАМН**

**Мун С.А., Глушков А.Н., Штернис Т.А.,
Ларин С.А., Максимов С.А.**

РЕГРЕССИОННЫЙ АНАЛИЗ В МЕДИКО-БИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

Методические рекомендации

**Кемерово
2012**

ŋqdqɏeɎ bɆ dədɆbɎɜə d ɕɔɑbɔɏ ɈqddɎqɏè

УТВЕРЖДАЮ

Директор
ФГБУН ИЭЧ СО РАН
профессор, д.м.н.

Глушков А.Н.
«23» августа 2012 г.

УТВЕРЖДАЮ

Ректор ГБОУ ВПО
КемГМА
профессор, д.м.н.

Ивойлов В.М.
«__» _____ 2012 г.

УТВЕРЖДАЮ

Начальник департамента
охраны
здоровья населения
Кемеровской области

Цой В.К.
«__» _____ 2012 г.

**Мун С.А., Глушков А.Н., Штернис Т.А.,
Ларин С.А., Максимов С.А.**

РЕГРЕССИОННЫЙ АНАЛИЗ В МЕДИКО-БИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

Методические рекомендации

**Кемерово
2012**

УДК [614.2:311](075.8)
ББК 51.1 (2)
В М904

Мун С.А.

В М904 Регрессионный анализ в медико-биологических исследованиях: методические рекомендации / С.А. Мун, А.Н. Глушков, Т.А. Штернис, С.А. Ларин, С.А. Максимов; ГБОУ ВПО КемГМА Минздравсоцразвития России. – Кемерово: КемГМА, 2012. – 115 с.

В методических рекомендациях представлены современные подходы к организации и проведению корреляционно-регрессионного анализа на примере парной линейной регрессии, а также нелинейных форм связи на примере степенной, показательной и гиперболической моделей.

В примерах подробно продемонстрировано пошаговое решение проверки значимости параметров и качества уравнения регрессии, выполнений условий Гаусса-Маркова, как в программе Microsoft® Excel®, Statsoft STATISTICA 6.0, так и с использованием математических формул. Уделено внимание интерпретации результатов и выводов.

Настоящие методические рекомендации предназначены для врачей-специалистов, аспирантов, ординаторов и интернов, студентов медицинского вуза.

УДК [614.2:311](075.8)
ББК 51.1 (2)

Рецензенты:

Главный областной специалист по научной работе медицинской службы КО, заместитель директора по научной работе ФГБУ НИИ Комплексных проблем сердечно-сосудистых заболеваний СО РАМН д.м.н., профессор Г.В. Артамонова

Директор ФГБУ НИИ Комплексных проблем гигиены и профессиональных заболеваний СО РАМН д.м.н., профессор В.В. Захаренков

Рассмотрено и рекомендовано к печати заседанием Ученого совета ИЭЧ СО РАН – 28.08.2012 г.;
Ученого совета КемГМА – 27.09.2012 г.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В современном обществе нет ни одной сферы человеческой деятельности, где бы ни применялась статистика, будь то экономика, экология, медицина, естественные науки, политология, социология, психология и т.д. С помощью статистики осуществляется научная обработка, обобщение и анализ информации, характеризующей развитие экономики страны, здравоохранения, политики, культуры и уровня жизни населения. Статистика позволяет выявить взаимосвязи (закономерности), изучить динамику развития, провести анализ для получения обоснованных выводов и принятия правильных решений, которые могут быть применены на практике.

Большим шагом в развитии медицинской статистической науки явилось применение математических методов и широкое использование компьютерной техники в анализе медико-биологических явлений.

Статистика, как любая наука, требует определения предмета исследования. Предметом статистики выступают размеры и количественные соотношения качественно определенных медико-биологических явлений, закономерности их взаимосвязей и развития в конкретных условиях места и времени. Свой предмет статистика изучает методом обобщающих показателей.

Для изучения предмета статистики разработаны и применяются специфические приемы, совокупность которых образует методологию статистики (методы массовых наблюдений, группировок, обобщающих показателей, динамических рядов, индексный метод и др.). Применение в статистике конкретных методов предопределяется поставленными задачами и зависит от характера исходной информации.

Исследование связей в условиях массового наблюдения и действия случайных факторов осуществляется, как правило, с помощью медико-статистических моделей. В широком смысле модель - это аналог, условный образ (изображение, описание, схема, чертеж и т.п.) какого-либо объекта, процесса или события, приближенно воссоздающий «оригинал». Модель представляет собой логическое или математическое описание компонентов и функций, отображающих существенные свойства моделируемого объекта или процесса, дает возможность установить основные закономерности изменения оригинала. В модели оперируют

ђqqbqеҮ bӖ dеqӖbҮзе d езqбсҮ Jqqbҭqуѐ

показателями, исчисленными для качественно однородных массовых явлений (совокупностей). Выражение моделей в виде функциональных уравнений используют для расчета средних значений моделируемого показателя по набору заданных величин и для выявления степени влияния на него отдельных факторов.

По количеству включаемых факторов модели могут быть однофакторными и многофакторными (два и более факторов).

В зависимости от познавательной цели статистические модели подразделяются на структурные, динамические и модели связи.

Наиболее разработанной в теории статистики является методология так называемой парной корреляции, рассматривающая влияние вариации факторного признака ***X*** на результативный признак ***Y*** и представляющая собой однофакторный корреляционный и регрессионный анализ.

СПИСОК ОСНОВНЫХ СОКРАЩЕНИЙ

Y	результативный признак (показатель)
$\hat{y}_i$	расчетное значение результативного признака, полученное по уравнению регрессии
X	факторный признак
ε	случайная ошибка или случайный член (ошибка измерений)
r_{xy}	коэффициент корреляции
a	параметр (коэффициент) уравнения регрессии, свободный член
b	параметр (коэффициент) уравнения регрессии
α	оценка параметра (коэффициента) уравнения регрессии
β	оценка параметра (коэффициента) уравнения регрессии
e_i	полагаемые значения (оценки) ошибок $\boldsymbol{\varepsilon_i}$
R^2	коэффициент детерминации
$R_{adj}^{\ 2}$	скорректированный коэффициент детерминации
n	число наблюдений
m	число параметров
S_e	стандартная ошибка
$Э_i$	средний коэффициент эластичности
$E_{отн\_i}$	средняя относительная ошибка аппроксимации
ДИ	доверительный интервал
МНК	метод наименьших квадратов
МВНК	метод взвешенного наименьшего квадрата
AR(1)	авторегрессионное преобразование первого порядка

1. КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫЙ АНАЛИЗ

Среди статистически взаимосвязанных признаков одни могут рассматриваться как определенные факторы, влияющие на изменение других, а вторые – как следствие, или результат изменения первых. Соответственно, первые – это факторные признаки, а вторые – результативные. Связь между двумя переменными ***X*** и ***Y*** является функциональной, если определенному значению переменной ***X*** соответствует строго определенное значение ***Y***. Это жестко детерминированная связь. Но существует и другая взаимосвязь, при которой взаимно действуют многие факторы, неравномерно влияющие на изменение результативного признака. Такие связи являются *стохастическими* (вероятностными).

Корреляционная связь является частным случаем стохастической связи. Это соотношение, соответствие между средним значением результативного признака и признаками-факторами. При этом если рассматривается связь средней величины результативного показателя ***Y*** с одним признаком-фактором ***X***, корреляционная связь называется «парной», а если факторных признаков два и более множественной. По характеру изменений ***Y***, ***X*** в парной корреляции различают прямую и обратную взаимосвязи. При прямой связи – с увеличением ***X*** возрастает и ***Y***, при обратной – уменьшается. По форме связи она делится на прямолинейные (линейные) и криволинейные (нелинейные).

Изучение корреляционных связей сводится к решению следующих задач:

1) выявление наличия или отсутствия корреляционной связи между изучаемыми признаками, эта задача может быть решена на основе параллельного сопоставления (сравнения) значений ***X*** и ***Y*** у ***n*** единиц совокупности, а также с помощью группировок и путем построения и анализа специальных корреляционных таблиц;

2) измерение тесноты связи между двумя и более признаками с помощью специальных коэффициентов (коэффициентов корреляции, $r_{x,y}$), и эта часть исследований называется «корреляционным анализом»;

3) определение уравнения регрессии – математической модели, в которой среднее значение результативного признака ***Y*** рассматривается как функция одной или нескольких переменных факторных признаков ***X*** и эта часть исследования носит название

2. СПЕЦИФИКАЦИЯ МОДЕЛИ (ДЛЯ МНОЖЕСТВЕННОЙ РЕГРЕССИИ)

При выборе факторных признаков для включения их в модель чаще всего руководствуются теоретическими представлениями о взаимосвязях факторов. Однако часто встречаются ситуации, когда имеется ***m*** число факторов, но нет априорной модели изучаемого фактора и не ясно, какие переменные можно включать в модель. В этом случае проводят спецификацию модели. Смысл понятия «спецификация модели»: это выбор объясняющих (существенных) и зависимых переменных и выбор функциональной зависимости.

Выбор объясняющих (существенных) переменных проводят методом пошагового отбора.

1. Из всего набора переменных отбирается (включается в модель) имеющая наибольший по модулю коэффициент корреляции с зависимой переменной ***y***.
2. На каждом последующем шаге в модель добавляется та из переменных, добавление которой максимально увеличивает скорректированный коэффициент детерминации R_{adj}^2 (только если соответствующая ***t***-статистика больше 1 или меньше -1).

Правильная функциональная зависимость (вид функции уравнения регрессии) должна отражать истинную зависимость между независимой ***x*** и зависимыми ***y*** переменными.

ŋqdqyeY bɆ dedɆbYɜe d ɛɑɒɔɑɣ ɟqdqYqɣè

3. АДЕКВАТНОСТЬ РЕГРЕССИОННОЙ МОДЕЛИ

Кроме проверки значимости параметров и качества уравнения регрессии в целом необходима проверка выполнений условий Гаусса-Маркова, обеспечивающих несмещенность и эффективность оценок параметров регрессии.

Оценка параметров регрессии является несмещенной, если математическое ожидание оценки равняется соответствующей характеристике генеральной совокупности. А оценка параметров регрессии будет эффективной, если она является надежной (точной) с определенным уровнем значимости (*p-level*), и чем он меньше, тем меньше вероятность ошибки (функция плотности вероятности распределения как можно более сжата вокруг истинного значения, т.е. дисперсия данной оценки минимальна).

Таким образом, если параметры и качество уравнения регрессии показали значимость уравнения линейной регрессии и были выполнены все условия Гаусса-Маркова, то такая модель будет считаться **адекватной**.

ПРОВЕРКА ВЫПОЛНЕНИЯ УСЛОВИЙ ГАУССА-МАРКОВА

Условия Гаусса-Маркова:

1. **Случайность остаточной компоненты** $Cov(X_i, e_i) = 0$**;**
2. **Равенство нулю математического ожидания средней величины остаточной компоненты** $M(\bar{e}) = 0$**;**
3. **Постоянства дисперсии случайного члена** e_i **во всех наблюдениях** $Var(e_i) = Const$**;**
4. **Независимость уровней ряда остатков** $Cov(e_i, e_j) = 0$**,** $i \neq j$**;**
5. **Соответствие ряда остатков закону распределения** $e_i \sim N(0, \sigma^2)$ **(не обязательное, но часто используемое условие).**

Достоинства:

1. Наиболее простой метод выбора значений ***a*** и ***b***, чтобы остатки были минимальными;

2. При выполнении условий Гаусса-Маркова МНК-оценки будут наилучшими (наиболее эффективными) линейными (комбинации y_i) несмещёнными оценками параметров регрессии (***a*** и ***b***).

Недостатки: МНК-оценки являются эффективными линейными несмещёнными ТОЛЬКО при выполнении ВСЕХ условий Гаусса-Маркова, что на практике встречается редко.

4. СРЕДНЯЯ ОТНОСИТЕЛЬНАЯ ОШИБКА АППРОКСИМАЦИИ

Для оценки точности регрессионных моделей используется средняя относительная ошибка аппроксимации $E_{отн\_i}$, которая показывает среднее отклонение расчетных значений $\hat{y}_i$ от фактических y_i и рассчитывается по формуле:

$$E_{отн\_i} = \frac{1}{n} \times \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{y_i} \times 100\% = \frac{1}{n} \times \sum_{i=1}^{n} \frac{e_i}{y_i} \times 100\%$$

где $\frac{|e_i|}{Y_i}$ – среднее значение относительной погрешности остатков;

n – количество наблюдений.

Если средняя относительная ошибка аппроксимации $E_{отн\_i} < 8-10\%$, то модель считается точной; если $10\% < E_{отн\_i} < 15\%$ – модель считается удовлетворительной.

5. СРЕДНИЙ КОЭФФИЦИЕНТ ЭЛАСТИЧНОСТИ

Средний коэффициент эластичности показывает, на сколько процентов изменится зависимая переменная y_i от своей средней величины при изменении независимой переменной x_i на 1% от своего среднего значения. Формулы расчетов коэффициентов эластичности для наиболее часто используемых типов уравнений регрессий приведены в таблице:

Тип регрессии	Уравнение регрессии	Средний коэффициент эластичности
линейная	$y = a + b \times x + \varepsilon$	$Э_i = b \times \frac{\bar{x}_i}{\bar{y}_i}$
степенная	$y = a \times x^b$	$Э_i = b$
гиперболическая	$y = a + \frac{b}{x} + \varepsilon$	$Э_i = -\frac{b}{a \times \bar{x} + b}$
показательная	$y = a \times b^x$	$Э_i = \bar{x} \ln b$

6. ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ ДЛЯ ПРОГНОЗНОГО ЗНАЧЕНИЯ

Интервальный прогноз (*ДИ*) для среднего значения показателя $\hat{y}$ рассчитывается по формуле: $\hat{y} \pm t_\alpha \times S_{прогноз}$

где: t_α (табличные значения *t*-критерия Стьюдента для односторонней области при уровне значимости p=0,05); S_e – стандартная ошибка модели;

$$S_{прогноз} = S_e \times \sqrt{\frac{1}{n} + \frac{(x_{прогноз.} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} .$$

А соответствующий *ДИ* для прогнозов индивидуальных значений (точечный прогноз) $\hat{y}_i$ будет рассчитываться по формуле:

$$\hat{y}_i \pm t_\alpha \times S_{прогноз}$$

где: $S_{прогноз} = S_e \times \sqrt{1 + \frac{1}{n} + \frac{(x_{прогноз.} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$

7. НЕЛИНЕЙНЫЕ ФОРМЫ ЗАВИСИМОСТИ

Использование линейной зависимости для описания данных наблюдений часто оказывается недостаточным. Необходимо использовать и нелинейные формы зависимостей, которые путем замены переменных можно преобразовать в линейный вид. Из нелинейных моделей чаще всего используются гиперболическая, степенная и показательная. Более подробно они описаны в разделе **ПРИМЕРЫ ПАРНОЙ РЕГРЕССИИ.**

ПРИМЕРЫ ПАРНОЙ РЕГРЕССИИ

ПРИМЕР 1.

Изучена еженедельная заболеваемость острыми респираторными инфекциями на территории Н. в зимний период (с декабря по февраль). Установлена корреляционная зависимость между средней еженедельной температурой в зимний период (***X***) и количеством острых респираторных заболеваний (ОРЗ) (***Y***).

Неделя	Кол-во ОРЗ	Температура воздуха, - t °С
1	30	20
2	31	21
3	33	22
4	34	23
5	34	21
6	36	25
7	38	25
8	39	29
9	38	28
10	36	23
11	28	20
12	34	22

Требуется:

1. Найти параметры уравнения линейной регрессии, дать интерпретацию коэффициента регрессии.

2. Вычислить остатки; найти остаточную сумму квадратов; оценить дисперсию остатков S_e^2; построить график остатков.

3. Проверить выполнение предпосылок МНК.

4. Осуществить проверку значимости параметров уравнения регрессии с помощью t-критерия Стьюдента (p=0,05). Дать интервальную оценку параметрам регрессии.

5. Вычислить коэффициент детерминации, проверить значимость уравнения регрессии с помощью F-критерия Фишера (p=0,05). Сделать вывод о качестве модели.

6. Найти коэффициент эластичности и среднюю относительную ошибку аппроксимации линейной регрессии.

7. Составить уравнения нелинейной регрессии:

- гиперболическую;
- степенную;
- показательную.

Найти коэффициенты детерминации, коэффициенты эластичности и средние относительные ошибки аппроксимации.

8. Сравнить модели по всем характеристикам и сделать вывод.

9. Осуществить прогнозирование значения показателя Y при уровне значимости p=0,05, если прогнозное значение фактора X составляет 80% от его максимального значения.

Вариант 1. С использованием математических формул
Вариант 2. В программе Excel
Вариант 3. С использованием программы STATISTICA 6.0

Решение задачи

Уравнение линейной модели парной регрессии: $\hat{y} = a + b \times x + \varepsilon$

1. Найти параметры уравнения линейной регрессии, дать интерпретацию параметра регрессии.

Вариант 1.

Для нахождения параметров уравнения линейной регрессии (***a***, ***b***) решим систему нормальных уравнений:

$$\begin{cases} a \times n + b\sum X_i = \sum Y_i \\ a\sum X_i + b\sum X_i^{\,2} = \sum X_i Y_i \end{cases}$$

Разделив обе части на ***n***, получим систему нормальных уравнений в виде:

$$\begin{cases} a + b \times \overline{X} = \overline{Y} \\ a \times \overline{X} + b \times \overline{X^2} = \overline{XY} \end{cases}$$

Решение этой системы даем нам найти параметры ***b*** и ***a*** по формулам:

$$b = \frac{\overline{XY} - \overline{Y} \times \overline{X}}{\overline{X^2} - \overline{X}^2} \quad \text{и} \quad a = \overline{Y} - b \times \overline{X}.$$

Предварительно, в программе Excel найдем промежуточные результаты, где ***Y*** (результативный признак) – количество острых респираторных заболеваний (ОРЗ), ***X*** (факторный признак) – средняя еженедельная температура в зимний период (°C) (табл. 1.1):

Таблица 1.1

i	Y_i	X_i	X_i^2	$Y_i \times X_i$
1	30	20	400	600
2	31	21	441	651
3	33	22	484	726
4	34	23	529	782
5	34	21	441	714
6	36	25	625	900
7	38	25	625	950
8	39	29	841	1131
9	38	28	784	1064
10	36	23	529	828
11	28	20	400	560
12	34	22	484	748
СРЗНАЧ	**34,250**	**23,250**	**548,583**	**804,500**

Далее, найдем по формулам параметры ***b*** и ***a*** уравнения регрессии:

$$b = \frac{804{,}5 - 34{,}25 \times 23{,}25}{548{,}583 - 23{,}25^2} = 1{,}0208216 = 1{,}0208;$$

$$a = 34{,}25 - 1{,}0208216 \times 23{,}25 = 10{,}5169$$

Вариант 2. В программе Excel строим таблицу (табл. 1.2), где ***Y*** (результативный признак) – количество острых респираторных заболеваний (ОРЗ), ***X*** (факторный признак) – средняя минусовая еженедельная температура в зимний период (°С).

Таблица 1.2

i	Y_i	X_i
1	30	20
2	31	21
3	33	22
4	34	23
5	34	21
6	36	25
7	38	25
8	39	29
9	38	28
10	36	23
11	28	20
12	34	22
СРЗНАЧ	**34,25**	**23,25**
СУММА	**411,00**	**279,00**

Далее, выбираем вкладку **Сервис – Анализ данных – Регрессия**, подставляем данные для входного интервала ***Y*** и ***X*** и выбираем ***остатки*** (рис. 1.1).

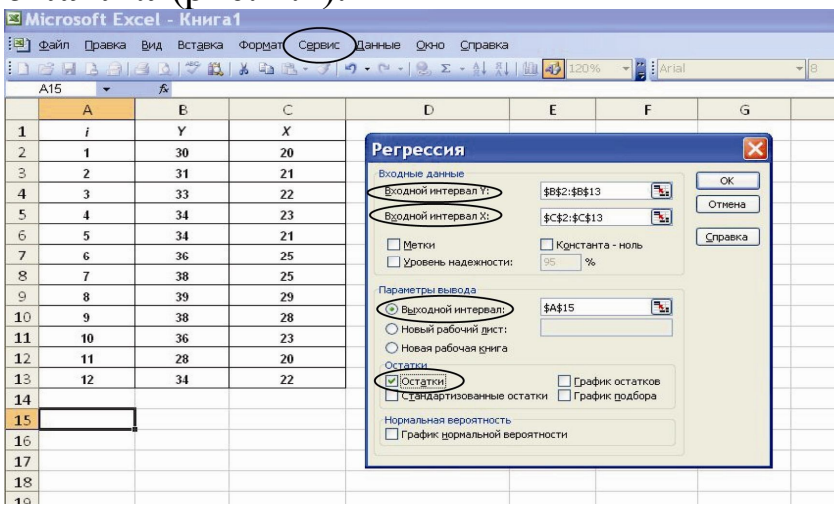


Рисунок 1.1.

ПРИМЕР 2

Представлены стандартизованные показатели (на 100000 населения) заболеваемости раком легкого с 1990 по 2005 гг. и выбросы загрязняющих веществ (ЗВ) в атмосферу (тыс. т) за период с 1985 по 2005 г. в г. N.

Требуется:

1. Установить зависимость влияния загрязнения атмосферного воздуха (***X***) на заболеваемость раком легкого (РЛ) (***Y***).

2. Найти параметры уравнения линейной регрессии. Осуществить проверку значимости параметров уравнения регрессии с помощью *t*-критерия Стьюдента (*p*=0,05). Дать интервальную оценку параметрам регрессии. Дать интерпретацию коэффициента регрессии.

3. Вычислить остатки; найти остаточную сумму квадратов; оценить дисперсию остатков S_e^2; построить график остатков.

4. Проверить выполнение предпосылок МНК.

5. Вычислить коэффициент детерминации, проверить значимость уравнения регрессии с помощью *F*-критерия Фишера (*p*=0,05), найти среднюю относительную ошибку аппроксимации и коэффициент эластичности.

6. Составить уравнения нелинейной регрессии:

- гиперболическую;
- степенную;
- показательную.

Найти коэффициенты детерминации, коэффициенты эластичности и средние относительные ошибки аппроксимации.

7. Сравнить модели по всем характеристикам и сделать вывод.

8. Осуществить прогнозирование значения показателя ***Y*** при уровне значимости *p*=0,05, если прогнозное значение фактора ***X*** составляет 80% от его максимального значения.

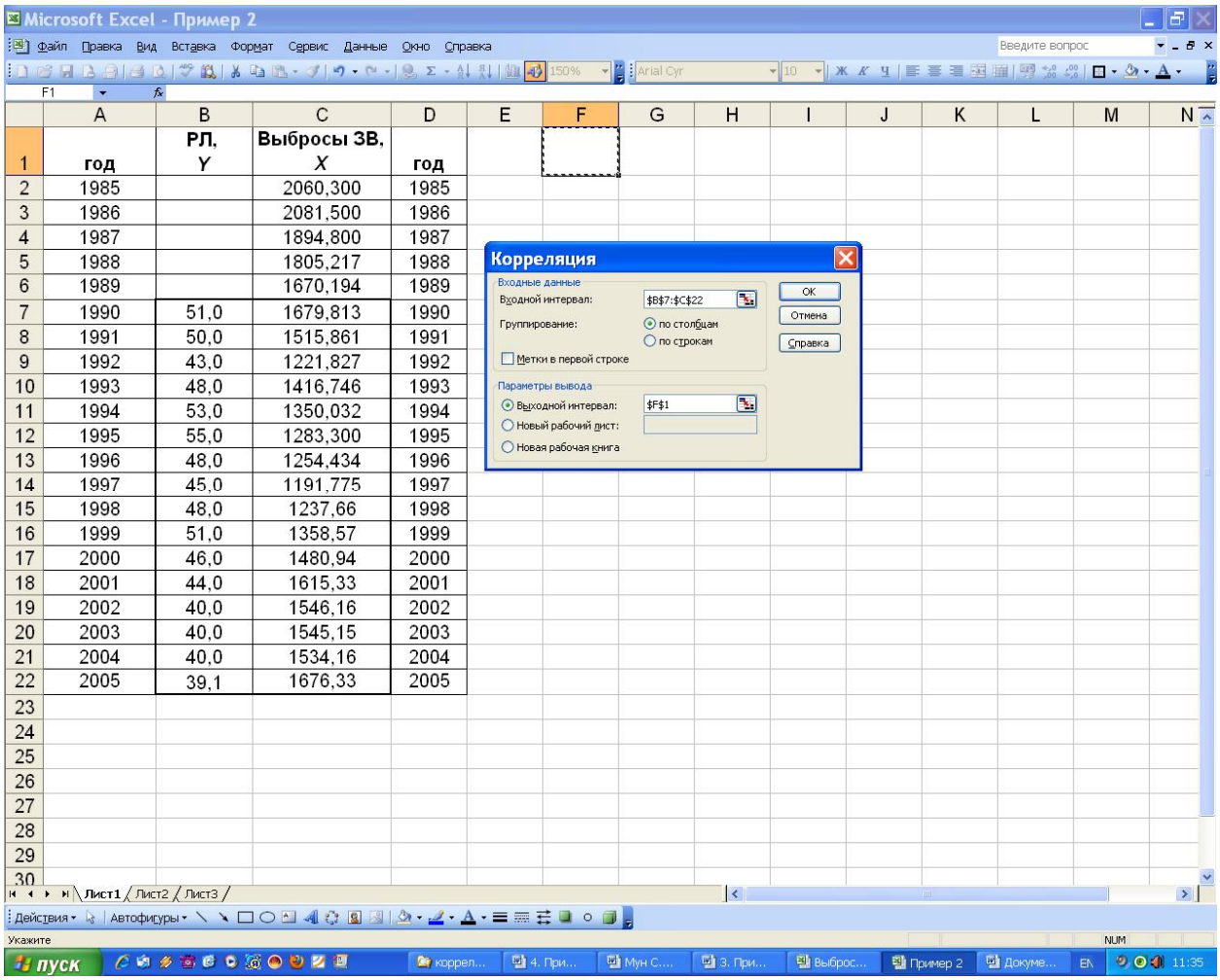
<u>**Решение задачи**</u>

1. Установить зависимость влияния загрязнения атмосферного воздуха (*X*) на заболеваемость раком легкого (*Y*).

При решении этого этапа следует учитывать длительность латентного периода возникновения рака, исходя из общих представлений о канцерогенезе. Иными словами, необходимо

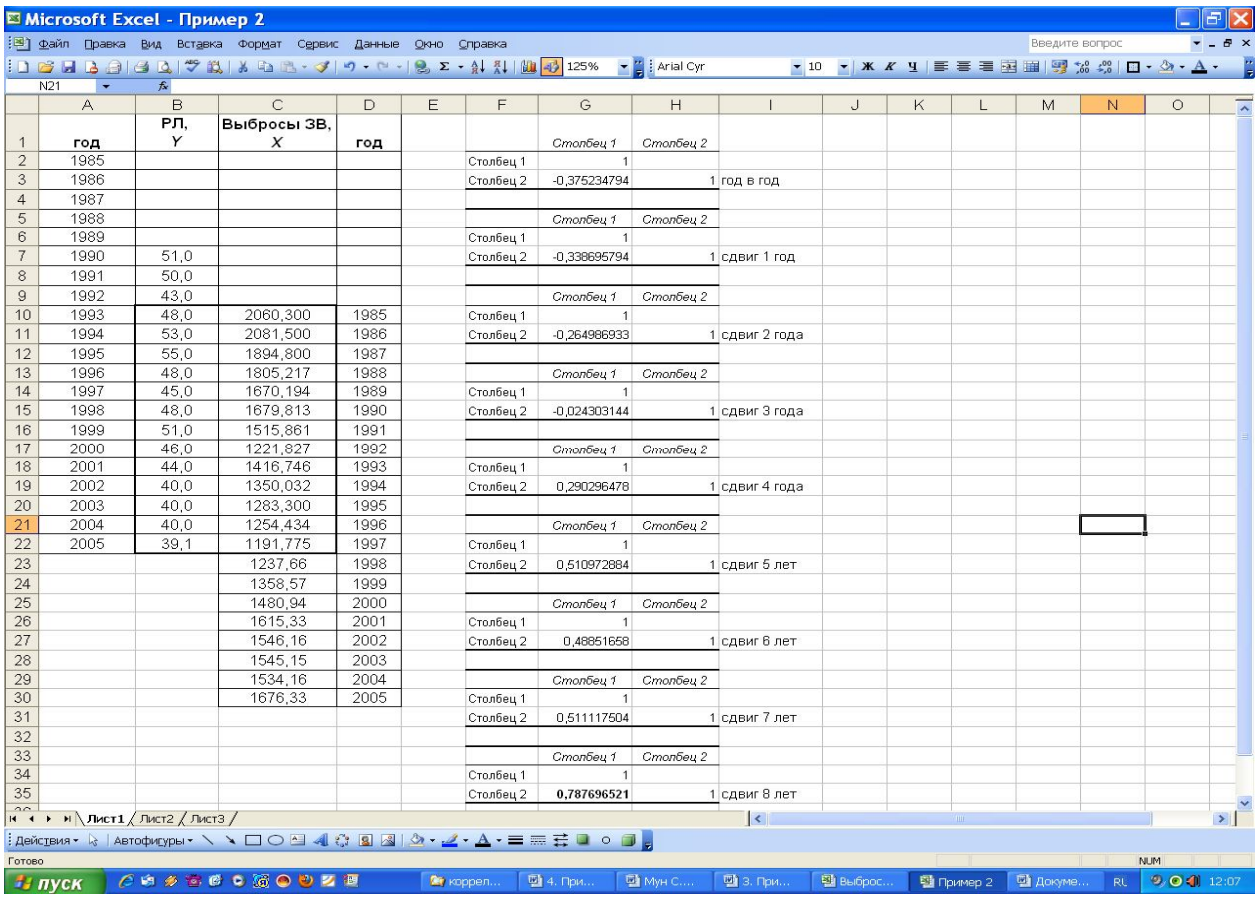
определить промежуток времени (***t***) между величиной выбросов ЗВ в атмосферу (***X***) и показателями заболеваемости РЛ (***Y***).

Для этого с помощью программы Excel (**Сервис – Анализ данных – Корреляция)** найдем тот сдвиг во времени, которому будет соответствовать статистически значимый коэффициент корреляции (рис. 2.1.и 2.2).



	A	B	C	D
1	год	РЛ, Y	Выбросы ЗВ, X	год
2	1985		2060,300	1985
3	1986		2081,500	1986
4	1987		1894,800	1987
5	1988		1805,217	1988
6	1989		1670,194	1989
7	1990	51,0	1679,813	1990
8	1991	50,0	1515,861	1991
9	1992	43,0	1221,827	1992
10	1993	48,0	1416,746	1993
11	1994	53,0	1350,032	1994
12	1995	55,0	1283,300	1995
13	1996	48,0	1254,434	1996
14	1997	45,0	1191,775	1997
15	1998	48,0	1237,66	1998
16	1999	51,0	1358,57	1999
17	2000	46,0	1480,94	2000
18	2001	44,0	1615,33	2001
19	2002	40,0	1546,16	2002
20	2003	40,0	1545,15	2003
21	2004	40,0	1534,16	2004
22	2005	39,1	1676,33	2005

Рисунок 2.1.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	год	РЛ, Y	Выбросы ЗВ, X	год			Столбец 1	Столбец 2							
2	1985					Столбец 1	1								
3	1986					Столбец 2	-0,375234794	1	год в год						
4	1987														
5	1988						Столбец 1	Столбец 2							
6	1989					Столбец 1	1								
7	1990	51,0				Столбец 2	-0,338695794	1	сдвиг 1 год						
8	1991	50,0													
9	1992	43,0					Столбец 1	Столбец 2							
10	1993	48,0	2060,300	1985		Столбец 1	1								
11	1994	53,0	2081,500	1986		Столбец 2	-0,264986933	1	сдвиг 2 года						
12	1995	55,0	1894,800	1987											
13	1996	48,0	1805,217	1988			Столбец 1	Столбец 2							
14	1997	45,0	1670,194	1989		Столбец 1	1								
15	1998	48,0	1679,813	1990		Столбец 2	-0,024303144	1	сдвиг 3 года						
16	1999	51,0	1515,861	1991											
17	2000	46,0	1221,827	1992			Столбец 1	Столбец 2							
18	2001	44,0	1416,746	1993		Столбец 1	1								
19	2002	40,0	1350,032	1994		Столбец 2	0,290296478	1	сдвиг 4 года						
20	2003	40,0	1283,300	1995											
21	2004	40,0	1254,434	1996			Столбец 1	Столбец 2							
22	2005	39,1	1191,775	1997		Столбец 1	1								
23			1237,66	1998		Столбец 2	0,510972884	1	сдвиг 5 лет						
24			1358,57	1999											
25			1480,94	2000			Столбец 1	Столбец 2							
26			1615,33	2001		Столбец 1	1								
27			1546,16	2002		Столбец 2	0,48851658	1	сдвиг 6 лет						
28			1545,15	2003											
29			1534,16	2004			Столбец 1	Столбец 2							
30			1676,33	2005		Столбец 1	1								
31						Столбец 2	0,511117504	1	сдвиг 7 лет						
32															
33							Столбец 1	Столбец 2							
34						Столбец 1	1								
35						Столбец 2	0,787696521	1	сдвиг 8 лет						

Рисунок 2.2.

В данном примере коэффициент корреляции r=0,79. Значимость r проверим с помощью критического значения коэффициента корреляции Пирсона $r_{крит}$ (приложение), при уровне значимости α=0,05:

$r_{крит}$=0,55, при n-2, что меньше r=0,79, следовательно коэффициент корреляции статистически значим;

и по t-критерия Стьюдента: найдем $t_{рассч} = \sqrt{\frac{n-2}{1-r_{x,y}^2}}$ и сравним с $t_{крит}$ (приложение) при (n-2); α=0,05/2 (двусторонняя область). Отсюда $t_{рассч} = \sqrt{\frac{13-2}{1-(0,79)^2}} = 5,41$; $t_{крит}$=2,20, следовательно, $t_{расч} > t_{крит}$, что говорит о значимости коэффициента корреляции.

Таким образом, выявлена по шкале Чеддока прямая, высокая корреляционная связь между заболеваемостью РЛ и выбросами ЗВ в атмосферу в г. N с промежутком во времени 8 лет.

Далее составим таблицу для дальнейших расчетов, где Y_i – заболеваемость РЛ на 100000 населения, а X_i – выбросы ЗВ в атмосферу (тыс. т) (табл. 2.1).

Таблица 2.1

i	Y_i	X_i
1	48,0	2060,300
2	53,0	2081,500
3	55,0	1894,800
4	48,0	1805,217
5	45,0	1670,194
6	48,0	1679,813
7	51,0	1515,861
8	46,0	1221,827
9	44,0	1416,746
10	40,0	1350,032
11	40,0	1283,300
12	40,0	1254,434
13	39,1	1191,775

2. Найти параметры уравнения линейной регрессии. Осуществить проверку значимости параметров уравнения регрессии с помощью *t*-критерия Стьюдента (*p*=0,05). Дать интервальную оценку параметрам регрессии. Дать интерпретацию коэффициента регрессии.

Параметры уравнения линейной регрессии получили с помощью программы Excel (**Сервис – Анализ данных – Регрессия)**, подставив данные для входного интервала ***Y*** и ***X*** и выбрав ***остатки*** (рис. 2.3).

ПРИМЕР 3

У 25 женщин, преподавателей среднеобразовательной школы, проведено измерение систолического артериального давления (САД) (мм.рт.ст.). Получена корреляционная зависимость между стажем преподавательской работы (***X***) и уровнем артериального давления (***Y***).

i	Y_i	X_i
1	110	2
2	100	4
3	110	7
4	170	35
5	110	4
6	115	9
7	110	5
8	90	7
9	115	24
10	110	9
11	90	8
12	110	9
13	160	22
14	110	30
15	100	3
16	105	11
17	100	8
18	130	8
19	125	8
20	120	13
21	110	9
22	160	30
23	145	32
24	100	9
25	150	41

Требуется:

1. Найти параметры уравнения линейной регрессии. Осуществить проверку значимости параметров уравнения регрессии с помощью *t*-критерия Стьюдента (p=0,05). Дать интервальную оценку параметрам регрессии. Дать интерпретацию коэффициента регрессии.

2. Вычислить остатки; найти остаточную сумму квадратов; оценить дисперсию остатков S_e^2; построить график остатков.

3. Проверить выполнение предпосылок МНК.

4. Вычислить коэффициент детерминации, проверить значимость уравнения регрессии с помощью F-критерия Фишера (p=0,05), найти среднюю относительную ошибку аппроксимации и коэффициент эластичности. Сделать вывод о качестве модели.

5. Составить уравнения нелинейной регрессии:

- гиперболическую;
- степенную;
- показательную.

Найти коэффициенты детерминации, коэффициенты эластичности и средние относительные ошибки аппроксимации.

6. Сравнить модели по всем характеристикам и сделать вывод.

7. Осуществить прогнозирование значения показателя Y при уровне значимости p=0,05, если прогнозное значение фактора X составляет 80% от его максимального значения.

Решение задачи

1. Найти параметры уравнения линейной регрессии. Осуществить проверку значимости параметров уравнения регрессии с помощью *t*-критерия Стьюдента (*p*=0,05). Дать интервальную оценку параметрам регрессии. Дать интерпретацию коэффициента регрессии.

Параметры уравнения линейной регрессии получили с помощью программы Excel (**Сервис – Анализ данных – Регрессия**), подставив данные для входного интервала ***Y*** и ***X*** и выбрав ***остатки*** (рис.3.1).

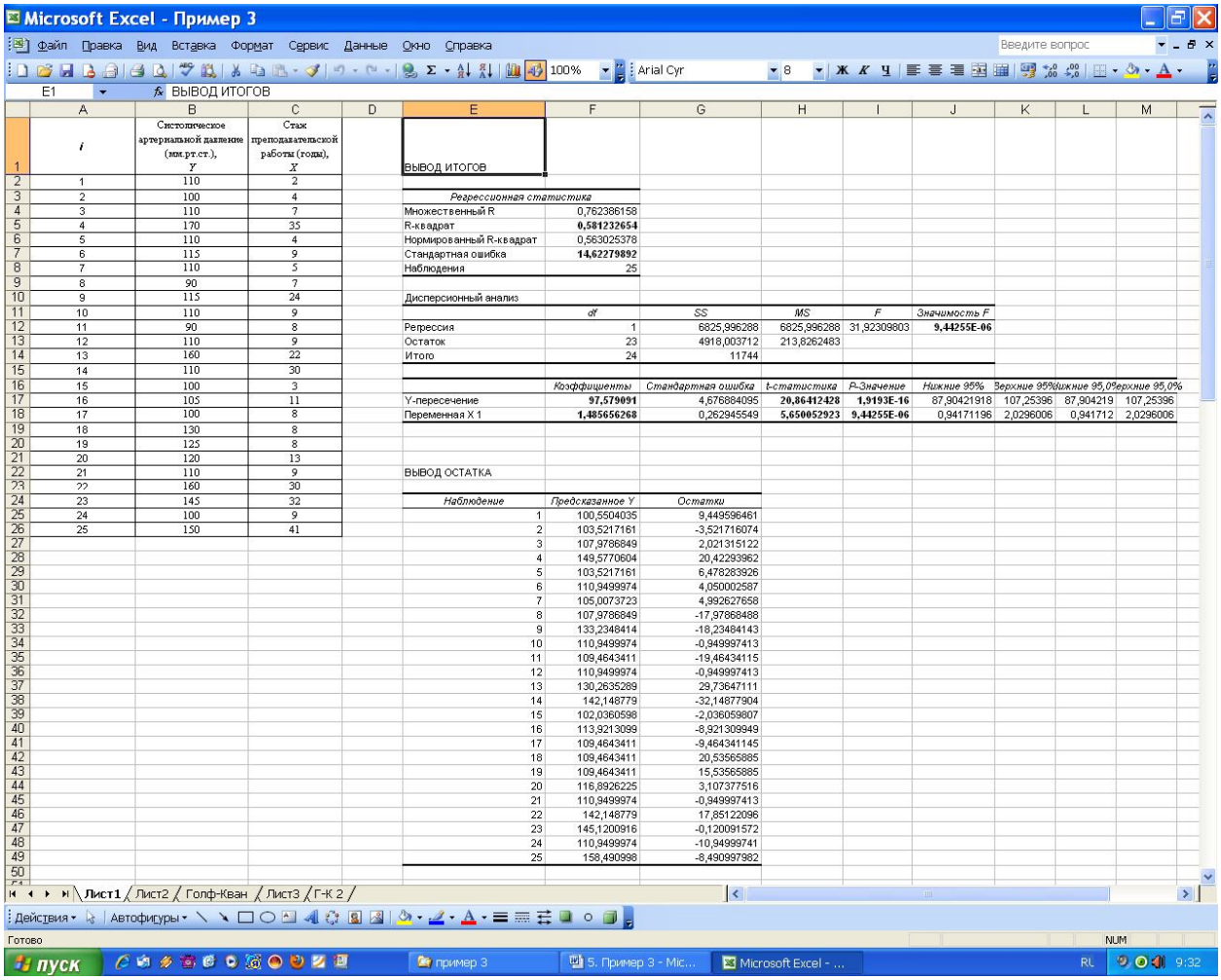


Рисунок 3.1.

Уравнение регрессии имеет следующий вид: $\hat{y} = 97{,}579 + 1{,}486 \times x$

Параметры регрессии *a*=97,579 (*t*=20,864; *p*=1,919E-16), *b*=1,486 (*t*=5,650; *p*=9,443E-06) статистически значимы. Параметр *b* означает, что при увеличении стажа работы на 1 год уровень систолического АД увеличивается в среднем на 1,486 мм.рт.ст.

2. Вычислить остатки; найти остаточную сумму квадратов; оценить дисперсию остатков S_e^2; построить график остатков.

Остатки получили методом регрессионного анализа (рис. 3.1). Остаточная сумма квадратов $SS_{ост}$ и дисперсия остатков $S_e^2=MS$ составили 4918,004 и 213,826 соответственно.

График остатков строим, используя в Excel надстройку «Мастер диаграмм»: тип диаграммы – точечная, выбираем столбцы ***Наблюдение*** и ***Остатки*** из таблицы **ВЫВОД ОСТАТКА** (рис. 3.1), где по оси абсцисс – наблюдения; по оси ординат – остатки (рис. 3.2).

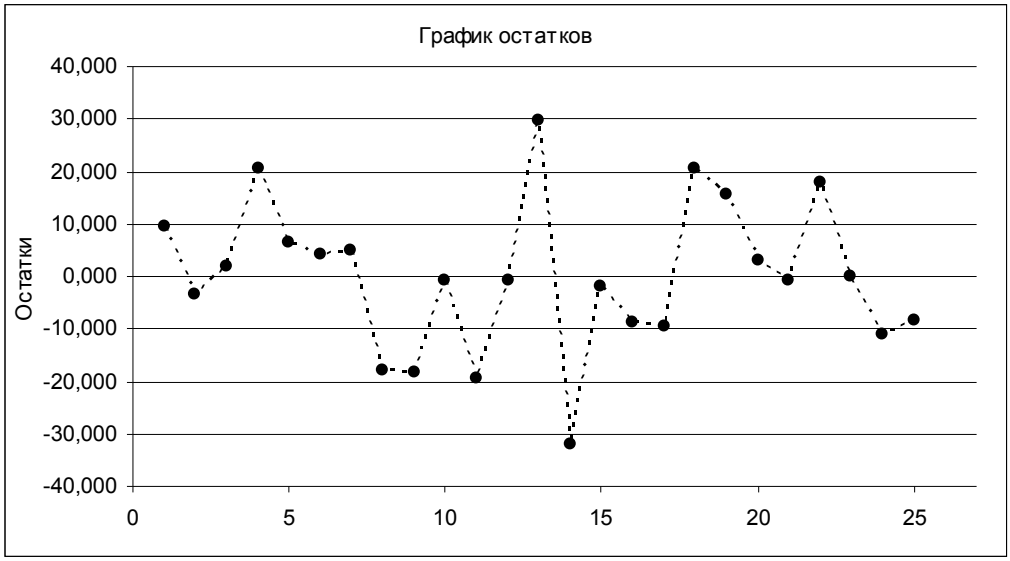


Рисунок 3.2.

3. Проверить выполнение предпосылок МНК.

Выполнение предпосылок МНК согласно условиям Гаусса-Маркова включают в себя проверку:

11) случайности остаточной компоненты $Cov(X_i,e_i)=0$ (критерий поворотных точек);
12) равенства нулю математического ожидания средней величины остаточной компоненты $M(\bar{e})=0$;
13) постоянства дисперсии случайного члена $\boldsymbol{e_i}$ во всех наблюдениях ($Var(e_i)=Const$) (критерий Голдфелда-Квандта, тест Спирмена);
14) независимости уровней ряда остатков $Cov(e_i, e_j)=0$, $j\neq i$ (критерий Дарбина-Уотсона);
15) соответствия ряда остатков закону распределения $e_i \sim N(0,\sigma^2)$ (*R/S*-критерий).

ПРИЛОЖЕНИЕ

Таблица критических значений коэффициентов корреляции Пирсона

Для уровня значимости α=0,05; α=0,01
Вероятность $p = \alpha$
где k – число степеней свободы

$k = n - 2$ \ α	0,05	0,01	$k = n - 2$ \ α	0,05	0,01
5	0,75	0,87	**27**	0,37	0,47
6	0,71	0,83	**28**	0,36	0,046
7	0,67	0,80	**29**	0,36	0,046
8	0,63	0,77	**30**	0,35	0,045
9	0,60	0,74	**35**	0,33	0,42
10	0,58	0,71	**40**	0,30	0,39
11	0,55	0,68	**45**	0,29	0,37
12	0,53	0,66	**50**	0,27	0,35
13	0,51	0,64	**60**	0,25	0,33
14	0,50	0,62	**70**	0,23	0,30
15	0,48	0,61	**80**	0,22	0,28
16	0,47	0,59	**90**	0,21	0,27
17	0,46	0,58	**100**	0,20	0,25
18	0,44	0,56	**125**	0,17	0,23
19	0,43	0,55	**150**	0,16	0,21
20	0,42	0,54	**200**	0,14	0,18
21	0,41	0,53	**300**	0,11	0,15
22	0,40	0,52	**400**	0,10	0,13
23	0,40	0,51	**500**	0,09	0,12
24	0,39	0,50	**700**	0,07	0,10
25	0,38	0,49	**900**	0,06	0,09
26	0,37	0,48	**1000**	0,06	0,09

Методические рекомендации

Мун Стелла Андреевна
Глушков Андрей Николаевич
Штернис Татьяна Александровна
Ларин Сергей Анатольевич
Максимов Сергей Алексеевич

РЕГРЕССИОННЫЙ АНАЛИЗ В МЕДИКО-БИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

Разработка макета – Мун С.А.
Ответственный редактор – Мун С.А.

Подписано в печать 27.09.12.
Тираж 100 экз. Формат $21\times30\,{}^{1}\!/_{2}$
Условных печатных листов 6,7