МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РФ
Федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования
«Московский авиационный институт
(национальный исследовательский университет)»

Кафедра «Моделирование систем и информационные технологии»

РЕГРЕССИОННЫЙ АНАЛИЗ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ
Однофакторный регрессионный анализ на базе программы "Stadia"

Методические указания к практическому занятию
по дисциплине "Математическая статистика"

Составители: Егорова Ю.Б.
Мамонов И.М.

МОСКВА 2020

ВВЕДЕНИЕ

Цель практического занятия - изучить порядок проведения однофакторного регрессионного анализа с помощью программы Stadia.

Основная задача регрессионного анализа заключается в нахождении математической зависимости (уравнения регрессии) $y=f(x)$ между исследуемыми факторами y и x.

В общем случае регрессионный анализ позволяет:

а) рассчитать коэффициенты регрессионной модели (коэффициенты регрессии) и проверить их значимость;

б) проверить адекватность регрессионной модели экспериментальным данным;

в) построить доверительные интервалы для уравнения регрессии;

г) выбрать из нескольких математических моделей ту, которая с большей точностью описывает экспериментальную зависимость;

д) использовать модель для прогнозирования.

1. ПОРЯДОК ВЫПОЛНЕНИЯ РАБОТЫ

1. Ввести **исходные данные** в электронную таблицу.
2. В блоке «Статистические методы» выбрать процедуру «**Простая регрессия**».
3. В меню "**Переменные регрессии**" выбрать для анализа переменные X и У из электронной таблицы.
4. В меню "**Регрессия**" необходимо выбрать регрессионную модель. В правой части меню выбора приведены формулы различных регрессионных моделей, в левой части - их названия.

5. Стандартная выдача **результатов регрессионного анализа** включает следующие характеристики:

5.1. **Регрессионную модель**, записанную в общем виде;

5.2. Таблицу, в которой приведены значения **коэффициентов регрессии**, стандартные ошибки вычисления каждого коэффициента, проверка значимости каждого коэффициента:

Коэффициент	a_0	a_1	a_i
Значение коэффициента			
Стандартная ошибка коэффициента			
Значимость *P* коэффициента			

Проверяется нулевая гипотеза "Коэффициент регрессии равен нулю", что означает его незначимость. Если $P>0{,}05$, то нулевая гипотеза принимается, следовательно, коэффициент незначим. Если $P<0{,}05$, то нулевая гипотеза отвергается и принимается альтернативная гипотеза "Коэффициент регрессии не равен нулю", что означает его значимость.

5.3. Таблицу **дисперсионного анализа**, предназначенную для анализа точности и адекватности модели:

Источники отклонений	Сумма квадратов отклонений	Степень свободы	Средн. сумма квадратов (дисперсия)
Регрессионные	$Q_{регр}$	$k_1 = p$	$S^2_{регр} = Q_{регр.}/k_1$
Остаточные	$Q_{ост}$	$k_2 = n\text{-}p\text{-}1$	$S^2_{ост} = Q_{ост}/k_2$
Общие	$Q_{общ} = Q^2_{регр.} + Q^2_{ост}$	$k_1 + k_2$	$S^2_{общ}$

5.4. Таблицу **проверки нулевой гипотезы** "Коэффициент корреляции равен нулю":

Множеств*R*	R^2	ПривR^2	Ст. ошиб.	*F*	Значим

Проверка **адекватности** регрессионной модели осуществляется в автоматическом режиме несколькими способами, но на экран монитора выводятся для наглядности результаты дисперсионного анализа и различные статистические характеристики:

- коэффициент корреляции *R*,
- коэффициент детерминации R^2,
- приведенный коэффициент детерминации R^2,
- стандартная ошибка уравнения регрессии,
- расчетное значение критерия Фишера *F*,
- уровень значимости *P* нулевой гипотезы.

Принятие нулевой гипотезы (при $P>0,05$) означает, что между исходными данными и выбранной регрессионной моделью нет соответствия, иными словами - модель неадекватно описывает экпериментальные данные. При $P<0,05$ - модель адекватна.

6. **Графическая выдача** результата регрессионного анализа содержит экспериментальные точки, регрессионную кривую, нижнюю и верхнюю границы доверительного интервала линии регрессии.

7. После выдачи результатов анализа следующим появляется меню "**Что дальше?**", предназначенное для выбора дальнейшего направления анализа из четырех пунктов: анализ остатков, прогнозирование, выбрать новую модель, закончить анализ.

7.1. **Анализ остатков** предназначен для нахождения доверительных интервалов уравнения регрессии. В ходе анализа остатков на экран монитора выводится таблица:

$X_{эксп}$	$У_{эксп}$	$У_{регр}$	остаток	Ст. остат.	Ст.ошиб.	Довер. Инт.

В таблице приводятся экспериментальные значения $X_{эксп}$ и $У_{эксп}$; значения $У_{регр}$, определенные по регрессионной модели; остаток $У_{эксп}$ - $У_{регр}$;

остаток в единицах стандартного отклонения; стандартная ошибка предсказания среднего значения; величина доверительного интервала при уровне надежности 0,95.

7.2. При выборе **прогноза** в правое поле меню дальнейшего анализа нужно ввести число точек прогноза, величину шага прогноза и нажать клавишу прогноза. На экран монитора выдается таблица:

$X_{прогн}$	$У_{прогн}$	Ст. ошиб.	Довер.инт

Таблица содержит прогнозируемое значение $X_{прогн}$; прогнозируемое значение $У_{прогн}$, определенное по регрессионной модели; стандартную ошибку прогноза, доверительный интервал прогноза при уровне надежности 0,95.

7.3. При выборе **новой модели** на экране монитора опять появляется меню "Регрессия" (см.п.4). Целесообразно рассчитать несколько моделей и среди моделей, адекватных экспериментальным данным, выбрать ту, для которой минимальна стандартная ошибка или максимален коэффициент корреляции.

7.4. Закончить анализ.

2. ОФОРМЛЕНИЕ ОТЧЕТА

Отчет должен содержать таблицу исходных данных, результаты процедур "Простая регрессия", "Анализ остатков", "Прогноз", два графика (линию регрессии с зоной доверительного интервала и график прогноза) и выводы.

ЗАДАНИЕ. Провести регрессионный анализ между сроком службы самолета (X, лет) и стоимостью его эксплуатации ($У$, млн руб.). Результаты измерений приведены в табл. 1.

I. Исходные данные:

Стоимость эксплуатации самолета в зависимости от срока службы:

X, лет	1	2	3	4	5
$У$, млн руб.	2	4	5	8	10

II. Результаты регрессионного анализа

ПРОСТАЯ РЕГРЕССИЯ. Файл: Стоимость эксплуатации самолета в зависимости от срока службы.

Переменные: x1, x2

Модель: линейная Y = a0+a1*x

Коэфф.	a0	a1
Значение	-0,2	2
Ст.ошиб.	0,5416	0,1633
Значим.	0,7323	0,0009

Источник	Сум.квадр.	Степ.св	Средн.квадр.
Регресс.	40	1	40
Остаточн	0,8	3	0,2667
Вся	40,8	4	

Множеств R	R^2	R^2прив	Ст.ошиб.	F	Значим
0,99015	0,98039	0,97386	0,5164	150	0

Гипотеза 1: <Регрессионная модель адекватна экспериментальным данным>

Хэксп	Yэксп	Yрегр	остаток	Ст.остат	Ст.ошиб	Довер.инт
1	2	1,8	0,2	0,4472	0,6532	2,052
2	4	3,8	0,2	0,4472	0,5888	1,85
3	5	5,8	-0,8	-1,789	0,5657	1,777
4	8	7,8	0,2	0,4472	0,5888	1,85
5	10	9,8	0,2	0,4472	0,6532	2,052

Хпрогн	Yпрогн	Ст.ошиб	Довер.инт
6	11,8	0,7483	2,351
7	13,8	0,8641	2,715
8	15,8	0,9933	3,121
9	17,8	1,131	3,555

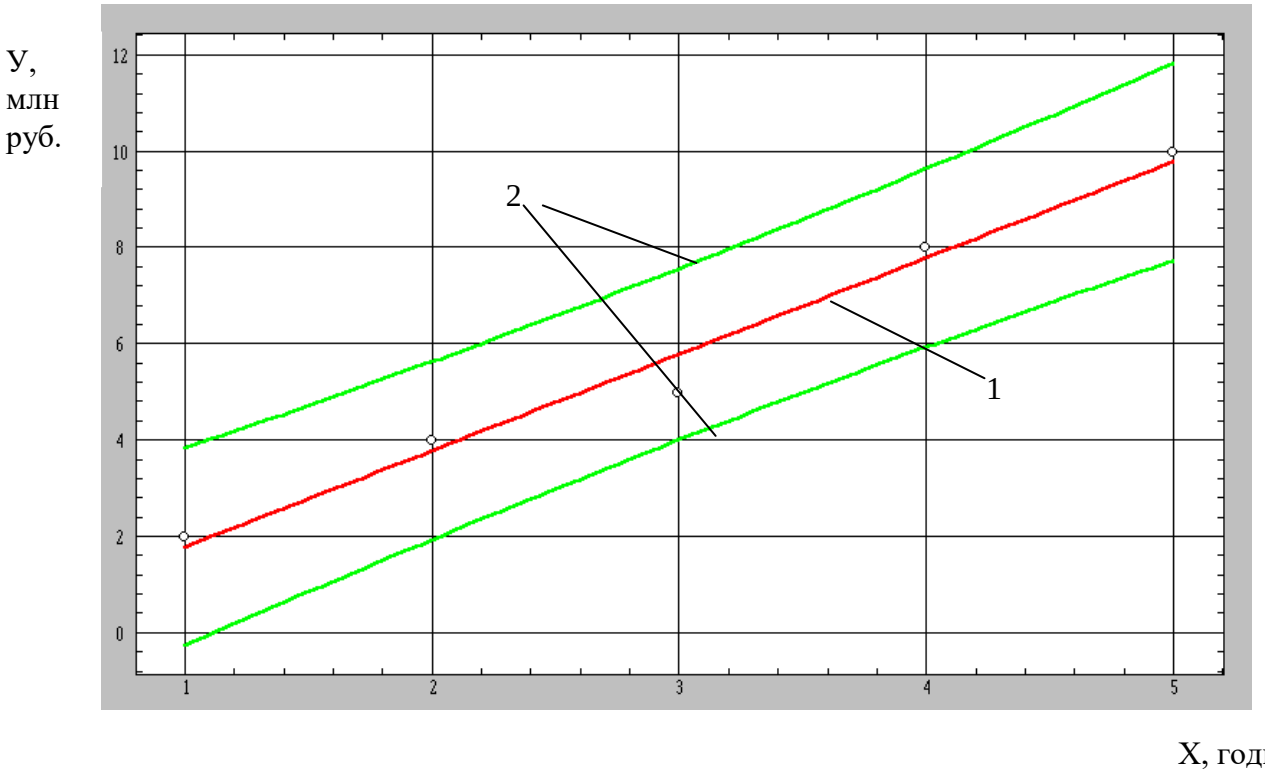

Рис. 1. Регрессионная зависимость стоимости эксплуатации самолета от срока службы: 1 - линия регрессии; 2 - границы доверительного интервала

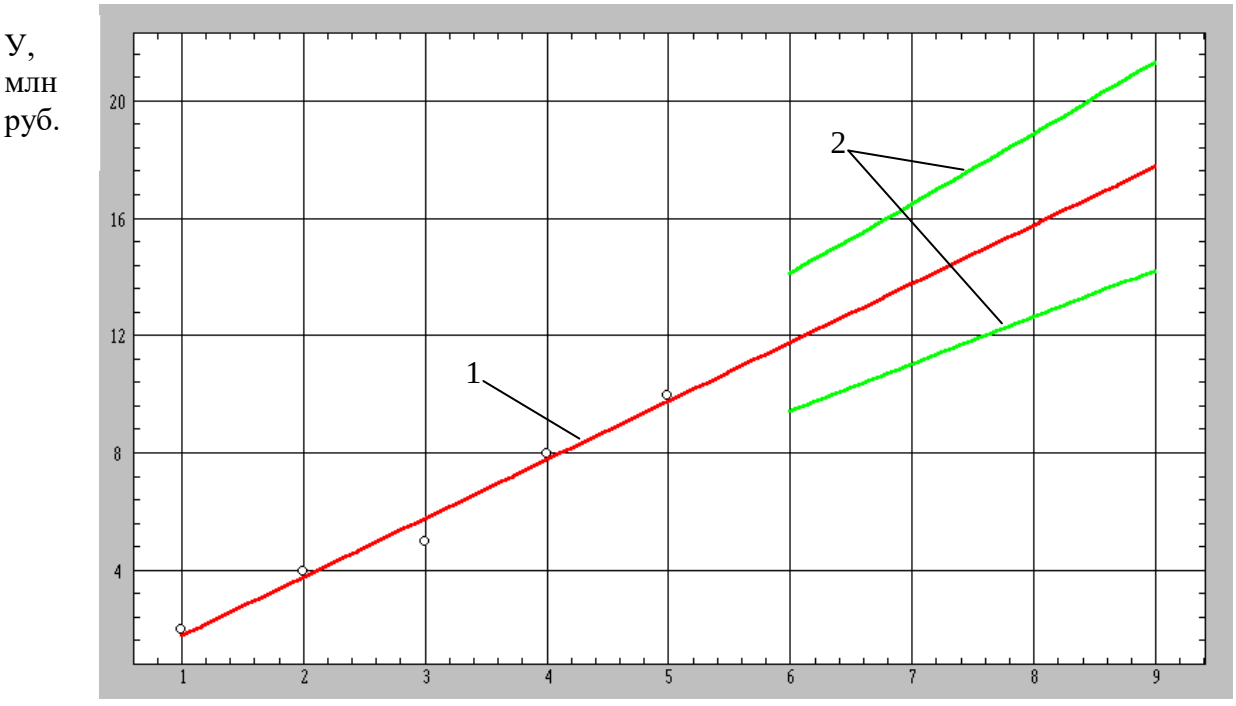

Рис. 2. Прогнозирование стоимости эксплуатации самолета от срока службы: 1 - линия регрессии; 2 - границы доверительного интервала прогноза индивидуальных значений

III. Выводы

1. Регрессионная зависимость между стоимостью эксплуатации самолета и сроком его службы имеет линейный вид:

$$У = -0{,}2 + 2X \quad \text{(млн руб.)}.$$

2. Коэффициент регрессии a_0=–0,2 незначим, так как P=0,7323> 0,05. Коэффициент регрессии a_1=2 значим, так как P=0,0009< 0,05.

3. Коэффициент корреляции $R=r$=0,99 свидетельствует об очень сильной прямой линейной связи между исследуемыми факторами.

4. Коэффициент детерминации R^2=0,98 говорит о том, что 98 % экспериментальных данных описаны данной регрессионной моделью.

5. Стандартная ошибка уравнения регрессии $S_{ост}$=0,5164 млн руб.

6. Регрессионная модель адекватна экспериментальным данным, так как значимость нулевой гипотезы H_0: "Коэффициент корреляции равен нулю" P=0,000… меньше критического значения 0,05.

7. Адекватность модели можно дополнительно проверить с помощью критерия Фишера F. Расчетное значение $F_{расч} = S^2_{регр}/S^2_{ост}$=40/0,2667=150. По таблицам распределения Фишера определяем табличное значение $F_{табл}$= 10,13 при степенях свободы k_1=1, k_2=3 и уровне значимости 0,05. Так как $F_{расч} > F_{табл}$, то уравнение адекватно описывает экспериментальные данные.

8. Прогноз: после 9 лет службы самолета прогнозируемая средняя стоимость его эксплуатации может составить 17,8 млн руб. с доверительным интервалом 3,555 млн руб., т.е. стоимость эксплуатации может колебаться от 14,245 до 21,355 млн руб.

3. ЗАДАНИЯ

Номер варианта задания определяется в соответствии с кодовой таблицей выбора вариантов (табл. 1) по начальным буквам фамилии и имени студента. Числовые данные к задаче приведены в табл. 2 и 3. В табл. 2 значения фактора x выбираются по начальной букве фамилии, в табл. 3 значения y – по начальной букве имени.

Таблица 1

Кодовая таблица выбора вариантов

Номер варианта	Начальные буквы	Номер варианта	Начальные буквы
1	А, Ж	9	П, Ф
2	Б, З	10	С, Ц
3	В, И	11	Ч, Ю
4	Г, К	12	У, Ш
5	Д, Р	13	Щ, Я
6	Е, М	14	Х, Э
7	Н, Т	15	Л
8	Г, О		

Например, Вас зовут Киров Александр. Ваш вариант по табл. 2 - 4, а по табл. 3 - 1. Вариант Вашего задания - 41.

Таблица 2

Числовые данные к задаче (*X*)

Вариант	Стоимость основных производственных фондов, млн руб.					
1	2,0	2,3	2,1	2,9	3,3	3,8
2	12,2	14,3	17,0	16,5	20,3	21
3	4,0	5,5	7,2	8,2	10,4	10,1
4	12,5	11,1	9,0	7,9	5,6	5,0
5	2,3	2,0	2,9	3,3	3,8	5,0
6	2,1	2,9	3,3	3,8	4,2	3,9
7	11,1	9,0	7,9	5,6	6,1	5,3
8	5,9	7,2	11,0	10,5	12,6	14,8
9	19,1	20,7	20,2	22,8	22,8	27,4
10	2,3	2,1	2,9	3,3	3,8	4,1
11	17,3	18,6	19,1	20,7	20,2	22,3
12	14,3	17,0	16,5	20,3	21,9	19,4
13	9,0	7,9	5,6	6,1	4,2	7,8
14	18,6	19,1	20,7	20,2	22,3	22,8
15	2,0	2,3	2,1	2,9	3,3	3,5

Таблица 3

Числовые данные к задаче (*У*)

Вариант	Суточная производительность, тыс. т					
1	17,3	18,6	19,1	20,7	20,2	22,3
2	112,0	104,3	99,6	95,4	83,0	80,1
3	18,6	19,1	20,7	20,2	22,3	25,4
4	24,0	29,4	34,2	30,6	35,2	36,0
5	19,1	20,7	20,2	22,3	22,8	18,4
6	111,0	104,3	99,6	95,4	83,0	92,3
7	34,2	30,6	35,2	40,7	43,5	44,2
8	29,3	34,2	30,6	35,2	40,7	44,5
9	64,5	70,2	79,3	74,6	81,4	93,0
10	23,9	24,7	22,4	25,1	27,0	29,2
11	33,8	30,6	37,8	40,2	41,5	44,3
12	104,3	99,6	95,4	83,0	86,4	81,5
13	68,3	64,5	70,2	79,3	82,6	101,4
14	29,4	34,2	30,6	35,2	40,9	43,5
15	57,8	68,3	64,5	70,2	79,3	83,6

4. КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Основная задача регрессионного анализа.
2. Этапы проведения регрессионного анализа.
3. Что характеризует коэффициент корреляции?
4. Как проверяется значимость коэффициента корреляции?
5. Как можно проверить адекватность регрессионной модели?
6. Что называется доверительным интервалом уравнения регрессии?
7. Как строится доверительный интервал уравнения регрессии?
8. Как можно оценить точность регрессионной модели?
9. Каким способом определяют коэффициенты регрессии?
10. Как проверяется значимость коэффициентов регрессии?
11. Как проводится прогнозирование?