Министерство науки и высшего образования РФ ФГАОУ ВПО
Национальный исследовательский технологический университет «МИСиС»

Институт компьютерных наук (ИКН)

Кафедра Автоматизированных систем управления (АСУ)

Курсовая работа
по дисциплине «Прикладной статистический анализ»

Выполнил:
студент группы БИВТ-20-3
Суриков А.С.

Проверил:
Гончаренко С.Н.

Москва, 2022

Оглавление

Введение

Статистический анализ данных широко используются во всех сферах деятельности: экономике, управлении, социологии, политологии, менеджменте, информатике, медицине, филологии, технических науках и др. Он позволяет найти взаимосвязи и выявить закономерности в изучаемых процессах и явлениях. В связи с постоянно растущим объемом данных возрастает потребность в использовании методов прикладного статистического анализа.

Прикладная статистика нацелена на решение реальных задач. Большую роль играет методологическая составляющая — как именно ставить задачи, какие предположения принять с целью дальнейшего математического изучения.

В настоящее время статистическая обработка данных проводится, как правило, с помощью соответствующих программных продуктов, таких как Statistica и MS Excel или Python.

В данной работе исследуются различные наборы данных, для каждого из которых выполняется определенный вид статистического анализа с соответствующей формой представления результатов.

Цель работы: исследовать различные наборы данных с применением различных видов статистического анализа.

Для достижения поставленной цели необходимо выполнить ряд задач:

- подобрать или подготовить наборы данных для анализа;
- построить основные графики по случайной выборке (график зависимости от параметра, гистограмму частот, диаграмму размахов, кривую распределений, графики скользящей средней);
- посчитать основные статистики по случайной выборке, рассчитать среднее хронологическое и средний квадрат ошибки (СКО);
- применить регрессионный анализ для случая парной линейной и нелинейной регрессии;
- применить регрессионный анализ для случая множественной линейной и нелинейной регрессии;
- выполнить кластерный анализ;
- выполнить факторный анализ;
- выполнить дисперсионный однофакторный анализ;
- выполнить дисперсионный двухфакторный анализ

1) Анализ закона распределения

В таблице 1 были собраны данные о времени запуска BackEnd компании “DataFrame” в секундах за период 01.11.2022 – 18.12.2022. Итого, размер выборки составил 48 элементов.

Таблица 1. Данные о времени запуска

Date	Server setup time
2022-11-01	62,51
2022-11-02	55,42
2022-11-03	56,33
2022-11-04	62,11
2022-11-05	62,75
2022-11-06	58,11
2022-11-07	60,22
2022-11-08	64,23
2022-11-09	62,78
2022-11-10	62,78
2022-11-11	63,91
2022-11-12	61,14
2022-11-13	62,77
2022-11-14	62,66
2022-11-15	62,68
2022-11-16	48,63
2022-11-17	60,14
2022-11-18	62,83
2022-11-19	44,26
2022-11-20	45,06
2022-11-21	50,1
2022-11-22	57,13
2022-11-23	54,36
2022-11-24	50,47
2022-11-25	52,73
2022-11-26	42,75
2022-11-27	47,59
2022-11-28	58,8
2022-11-29	61,77
2022-11-30	46,12
2022-12-01	61,54
2022-12-02	45,94
2022-12-03	49,66
2022-12-04	42,1
2022-12-05	44,08
2022-12-06	40,28
2022-12-07	49,3
2022-12-08	41,7
2022-12-09	56,11
2022-12-10	51,26
2022-12-11	51,05

2022-12-12	42,97
2022-12-13	50,36
2022-12-14	61,85
2022-12-15	54,83
2022-12-16	58,78
2022-12-17	47,67
2022-12-18	55,96

1) На рисунке 1 представлена динамика кол-ва свободных номеров в зависимости от даты.

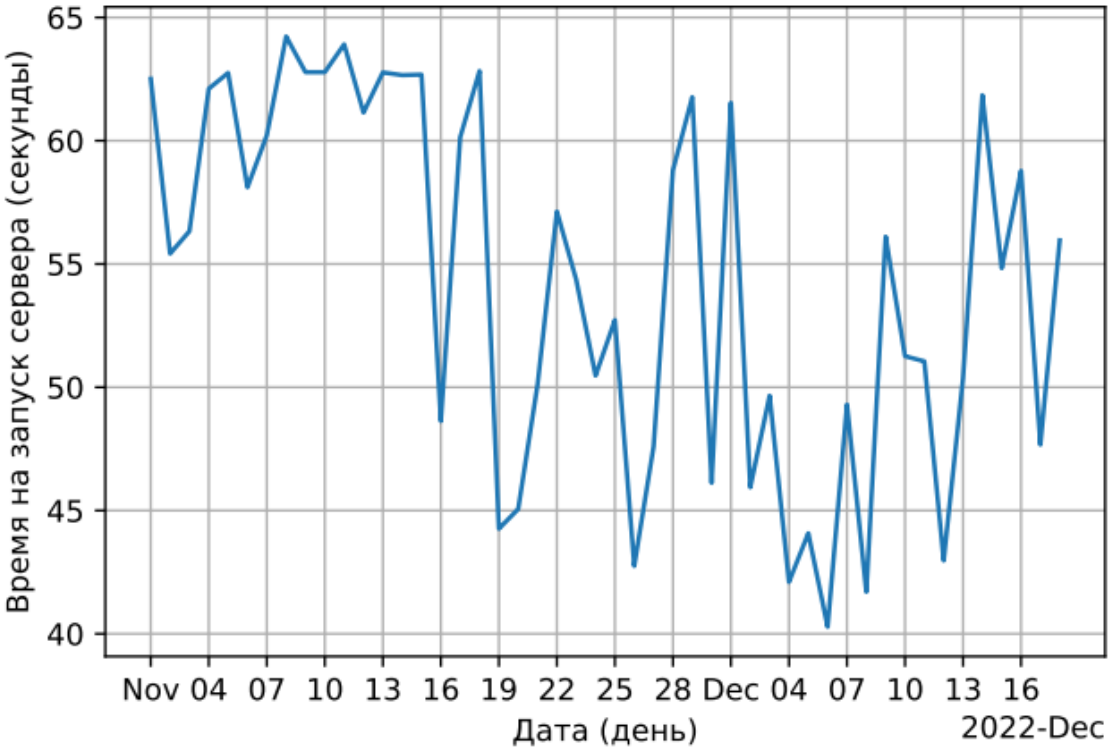

Рисунок 1. Динамика зависимости времени запуска от даты

2) В таблице 2 отображены рассчитанные основные статистики.

Таблица 2. Описательные статистики

Описательные статистики	Величина	Единица измерения
Кол-во наблюдений	48	
Среднее	54.21	секунды
Хронологическое среднее	55.51	секунды
Геометрическое среднее	53.53	секунды
Медиана	55.23	секунды
Мода	62	секунды
Минимум	40.28	секунды
Максимум	64.23	секунды
Нижний квантиль	48.21	секунды
Верхний квантиль	61.79	секунды
Размах	24.28	секунды
Дисперсия	55.4	(секунды)^2
Стандартное отклонение	7.44	секунды
Асимметрия	-1.282	
Эксцесс	-0.304	

3) По формуле Стерджесса, представленной на рисунке 2, определили количество интервалов, необходимых для графического представления набора статистических данных. На рисунке 3 представлена гистограмма относительных частот с кривой нормального распределения. Смещение значений вправо относительно нормального распределения подтверждает отрицательный коэффициент асимметрии. Смещение значений вверх относительно нормального распределения подтверждает положительный коэффициент эксцесса.

$$n = 1 + 3{,}322 \cdot \lg N$$

Рисунок 2. Формула Стерджесса

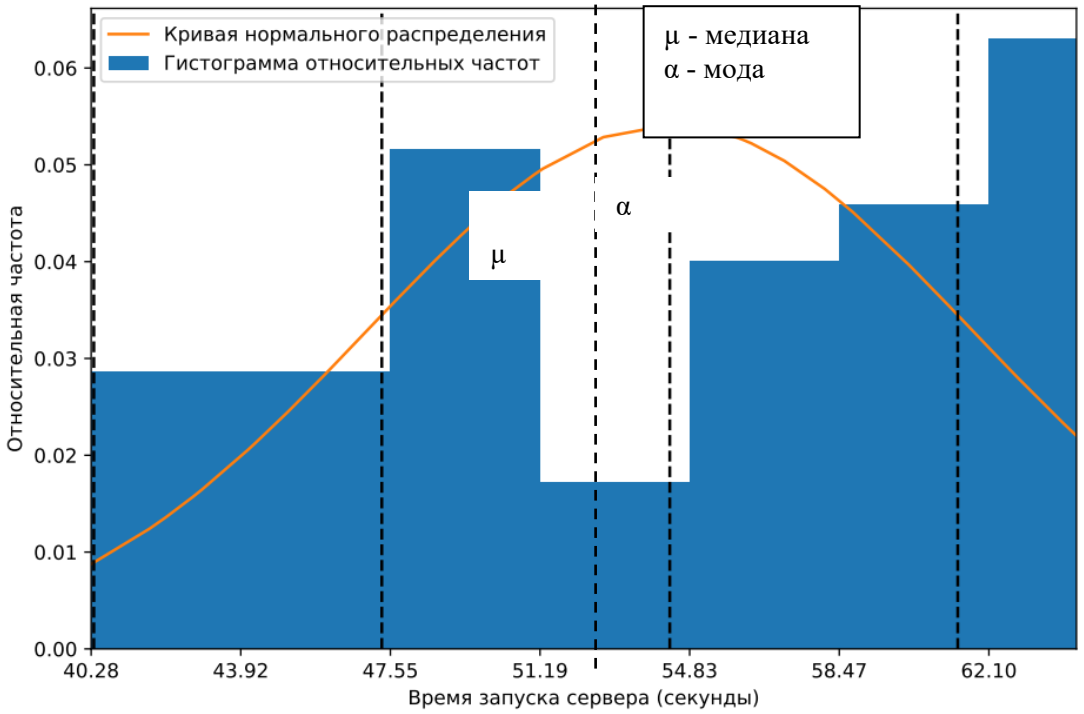

Рисунок 3. Гистограмма относительных частот

На рисунке 4 представлена кривая нормального распределения с коэффициентами асимметрии (А) и эксцессом (Ex) равными нулю.

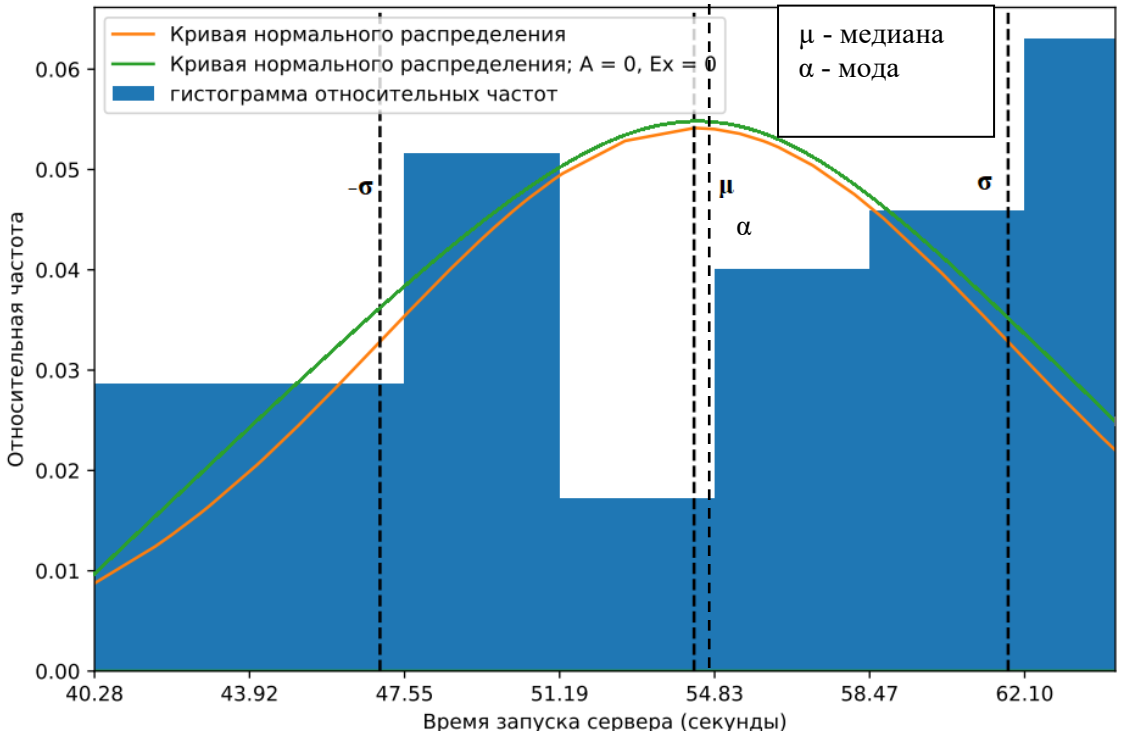

Рисунок 4. Кривая нормального распределения с коэффициентами асимметрии и эксцессом равными нулю

4) Рассчитаем критерий хи-квадрат по формуле представленной на рисунке 5. Эмпирические и теоретические частоты представлены в таблице 3. Полученное значение критерия хи-квадрат наблюдаемое сравним со значением хи-квадрат критическое, взятое для числа степеней свободы – 6 и уровня значимости а – 0.05. На основании выполнения неравенства 1.03 < 12.59 ($\chi^2_{набл} < \chi^2_{кр}$) делаем вывод, что контрольные данные подчиняются закону распределения Гаусса.

$$\chi^2_{набл} = \sum_{i=1}^{k} \frac{(f_{эмп} - f_{теор})^2}{f_{теор}}$$

$f_{эмп}$ - эмпирическая частота
$f_{теор}$ - теоретическая частота
k - количество разрядов

Рисунок 5. Формула вычисления критерия хи-квадрат

Таблица 3. Значения эмпирических и теоретических частот

Номер	Эмпирическая частота	Теоретическая частота
0	0,200485	0,095222
1	0,200485	0,191489
2	0,360874	0,301811
3	0,120291	0,37283
4	0,280679	0,360968
5	0,320777	0,273912
6	0,441068	0,162906
$\chi^2_{кр}$	12.59	
$\chi^2_{набл}$	0.54	

5) По периодам 3, 4, 5, 6 и 10 были получены значения скользящих средних и был рассчитан средний квадрат ошибки периодов (СКО), значения представлены в таблице 4. Зависимость СКО от периода скользящих средних представлена на графике на рисунке 6.

Таблица 4. Значения для скользящих средних и их СКО

Номер операции	mv3	mv4	mv5	mv6	mv10
1	0	0	0	0	0
2	58,09	59,09	0	0	0
3	57,95	59,15	59,82	59,54	0
4	60,4	59,82	58,94	59,16	0
5	60,99	60,8	59,9	60,62	60,72
6	60,36	61,33	61,48	61,7	60,86
7	60,85	61,34	61,62	61,81	61,44
8	62,41	62,5	61,62	62	62,08
9	63,26	63,42	62,78	62,51	62,14
…	63,16	62,65	62,97	62,94	62,13
40	62,61	62,65	62,68	62,67	61,18
41	62,61	62,62	62,65	62,66	61,17
42	62,19	62,31	62,63	60,3	61,03
43	62,7	59,18	59,58	59,67	59,18
44	57,99	58,53	59,38	59,95	57,41
45	57,15	58,57	59,39	56,87	56,03

46	57,2	53,96	55,71	53,93	55,63
47	55,74	53,07	52,18	51,84	54,78
48	50,72	50,56	52,48	53,25	53,57
СКО	17.48	15.84	14.73	14.03	10.74

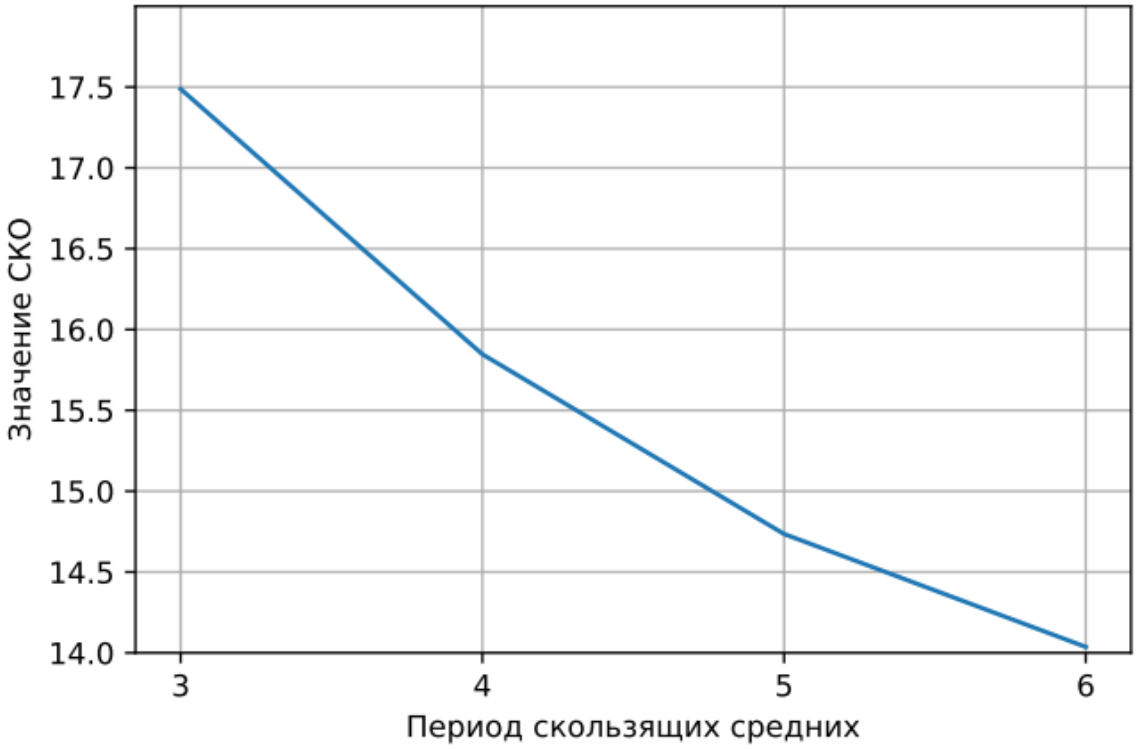

Рисунок 6. График зависимости СКО от периода скользящих средних

6) На рисунке 7 представлен график динамики кол-ва свободных номеров отеля, а также кривая скользящих средних с наименьшим средним квадратом ошибки (по периоду 10).

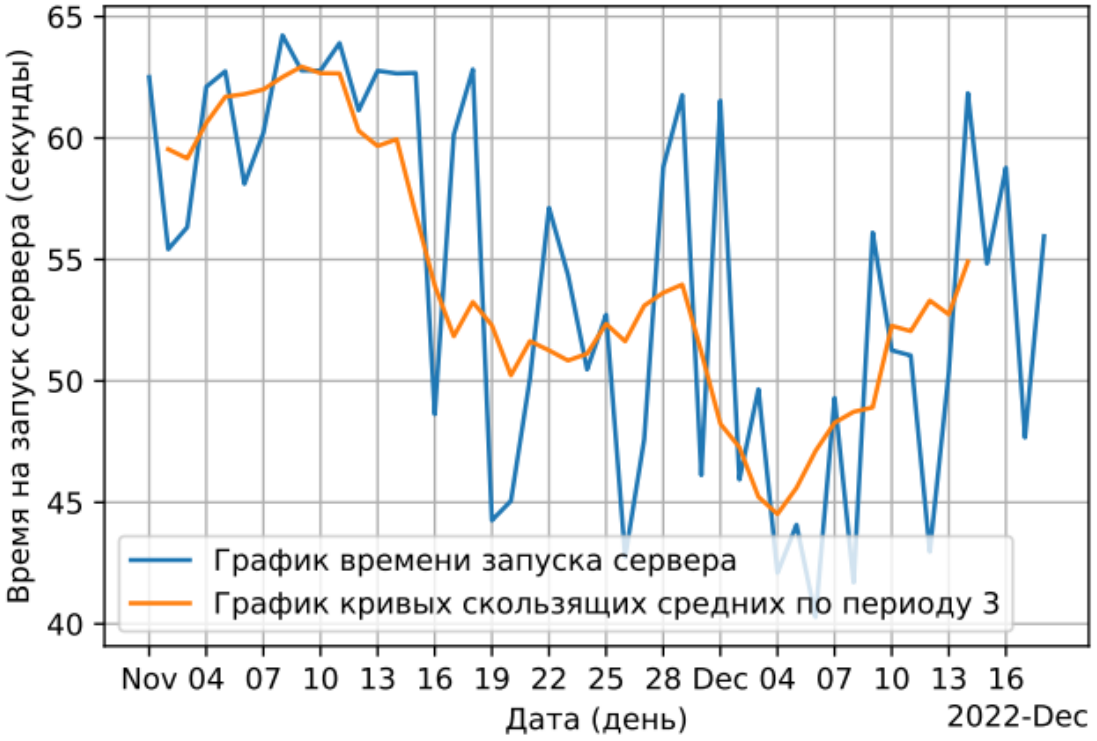

Рисунок 7. График с кривой скользящей средней

7) На рисунке 8 представлены график динамики кол-ва свободных номеров отеля с кривыми краткосрочной и долгосрочной средних (индикаторы), на основании которых были отмечены золотые пересечения, широко используемые в биржевом анализе.

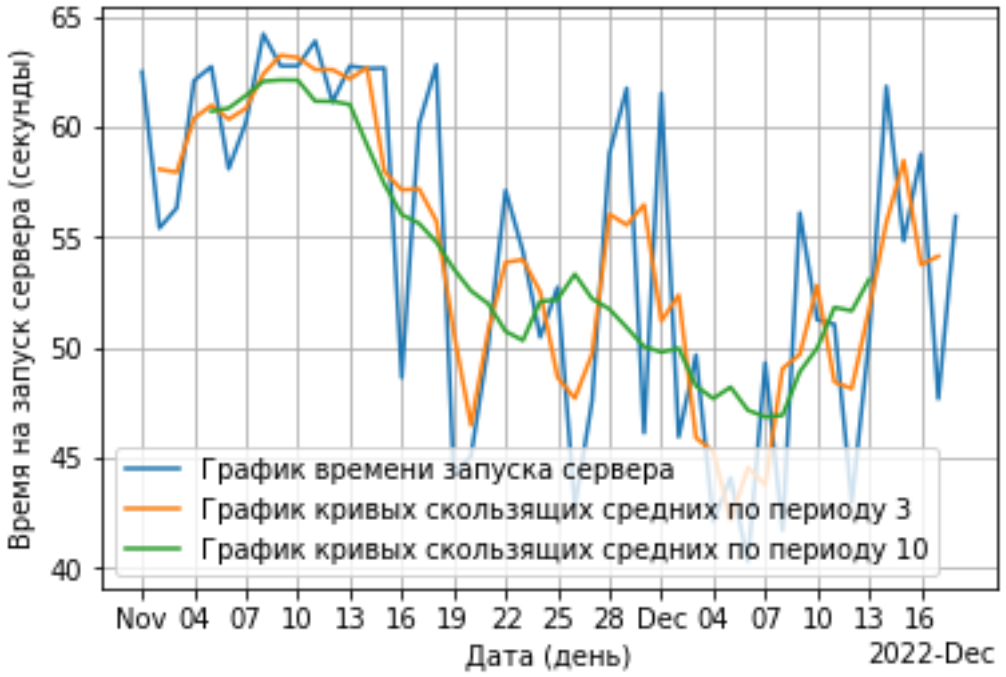

Рисунок 8. График динамики кол-ва свободных номеров отеля с кривыми с краткосрочной и долгосрочной средними

8) На рисунке 9 представлена диаграмма размаха. Такой вид диаграммы в удобной форме показывает медиану, нижний и верхний квартили, минимальное и максимальное значение выборки и выбросы.

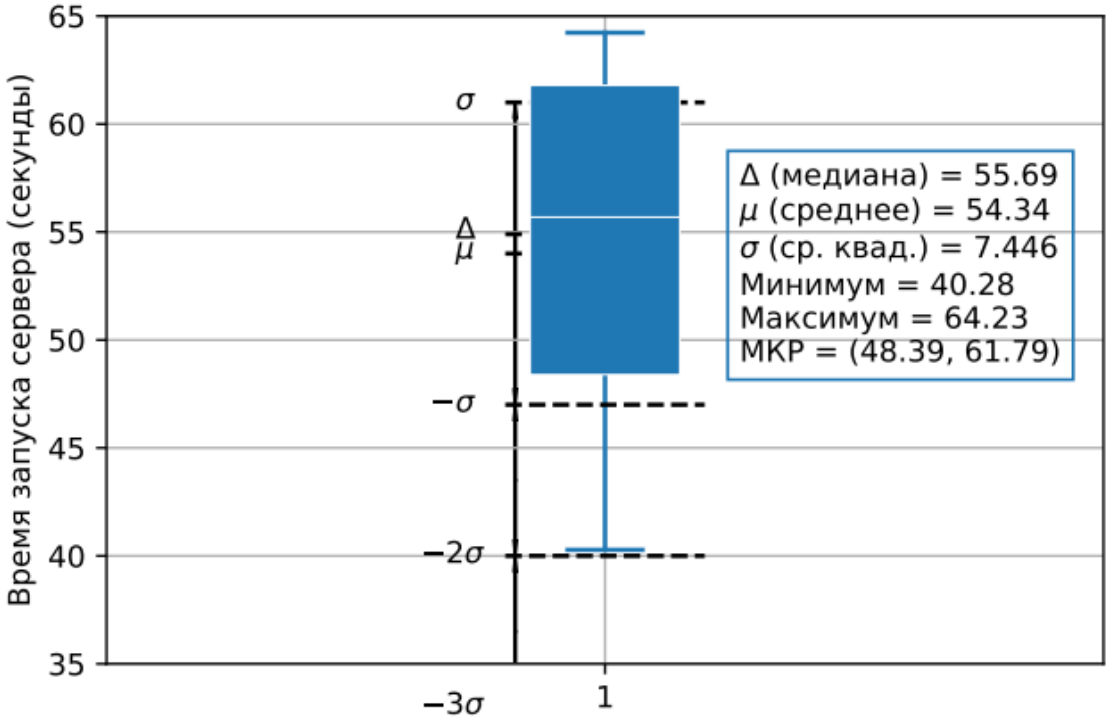

Рисунок 9. Диаграмма размаха

2) Регрессия

Были собраны данные о рейтинге лиги игры в многопользовательской стратегии “StarCraft 2” в зависимости от четырех числовых параметров. Целевая переменная $y(x_1, x_2, x_3, x_4)$ – APM (Action Per Minute) - количество нажатий мышкой в среднем в минуту за игру; x_1 – возраст игрока в годах, x_2 – количество часов в неделю, которые игрок тратит на игру в “StarCraft 2”, x_3 – общее количество часов игрока в “StarCraft 2”, x_4 – лига игрока: бронзовая, серебряная, золотая, платиновая, алмазная, мастер, гроссмейстерская и профессиональная лига с кодами 1-8 (навык игрока оцениваемый системой).

Часть 1. Линейная и нелинейная регрессия.

Для первой части лабораторной работы мы проанализируем только один фактор – возраст игрока, и попробуем выяснить зависимость APM от возраста играющего.

Предположим различные типы зависимостей: ax+b, x^2, x^3, ln(x), x^a, и для каждой из них построим линию тренда, а также посчитаем коэффициент корреляции и средний квадрат ошибки.

Табл.1. Таблица зависимостей АРМ от возраста

Номер	АРМ	Возраст (лет)
1	143,718	27
2	129,232	23
3	69,9612	30
4	107,602	19
5	122,891	32
6	44,457	27
7	46,9962	21
8	212,602	17
9	117,488	20
10	155,986	18
11	153,801	16
12	79,2948	26
13	67,4754	18
14	119,437	38
15	160,475	16
16	81,7722	17

17	50,8374	28
18	160,646	20
19	107,912	16
20	114,781	26
21	115,127	21
22	133,702	21
23	99,5088	18
24	83,9172	26
25	216,694	17
26	129,86	23
27	267,559	18
28	74,1174	25
29	101,68	25

Табл.2. Определение вида и формы зависимости

АРМ	Возраст	y = -0,0413x+27,315	y = -4,511ln(x)+43,556	y = -4E-05x2 - 0,0308x + 26,679	y = 3E-06x3 - 0,0015x2 + 0,1563x + 19,747	y = 58,576x^-0,211
143,72	27	21,38	21,15	23,29	29,04	20,53
129,23	23	21,98	21,63	23,53	27,84	21,00
69,96	30	24,43	24,39	24,77	25,39	23,90
107,60	19	22,87	22,45	23,94	26,67	21,83
122,89	32	22,24	21,85	23,65	27,44	21,22
44,46	27	25,48	26,44	25,41	24,26	26,30
47,00	21	25,37	26,19	25,34	24,40	26,00
212,60	17	18,53	19,38	22,39	42,83	18,91
117,49	20	22,46	22,06	23,75	27,14	21,43
155,99	18	20,87	20,78	23,09	30,40	20,18
153,80	16	20,96	20,84	23,12	30,13	20,24
79,29	26	24,04	23,83	24,55	25,70	23,28
67,48	18	24,53	24,56	24,83	25,31	24,09
119,44	38	22,38	21,98	23,71	27,24	21,35
160,48	16	20,69	20,65	23,02	31,00	20,06
81,77	17	23,94	23,69	24,49	25,78	23,13
50,84	28	25,22	25,83	25,24	24,60	25,57
160,65	20	20,68	20,64	23,02	31,02	20,06
107,91	16	22,86	22,44	23,94	26,69	21,81
114,78	26	22,57	22,16	23,80	27,00	21,53
115,13	21	22,56	22,15	23,80	27,02	21,52
133,70	21	21,79	21,47	23,45	28,17	20,85
99,51	18	23,21	22,80	24,11	26,36	22,19
83,92	26	23,85	23,57	24,45	25,85	23,00
216,69	17	18,37	19,29	22,35	44,23	18,83
129,86	23	21,95	21,60	23,52	27,89	20,98
267,56	18	16,26	18,34	22,02	69,11	18,01
74,12	25	24,25	24,13	24,67	25,53	23,61
101,68	25	23,12	22,71	24,06	26,44	22,09

Посмотрим на линию тренда каждой из зависимостей, а также рассчитаем коэффициент корреляции R^2 и составим модели уравнений регрессии.

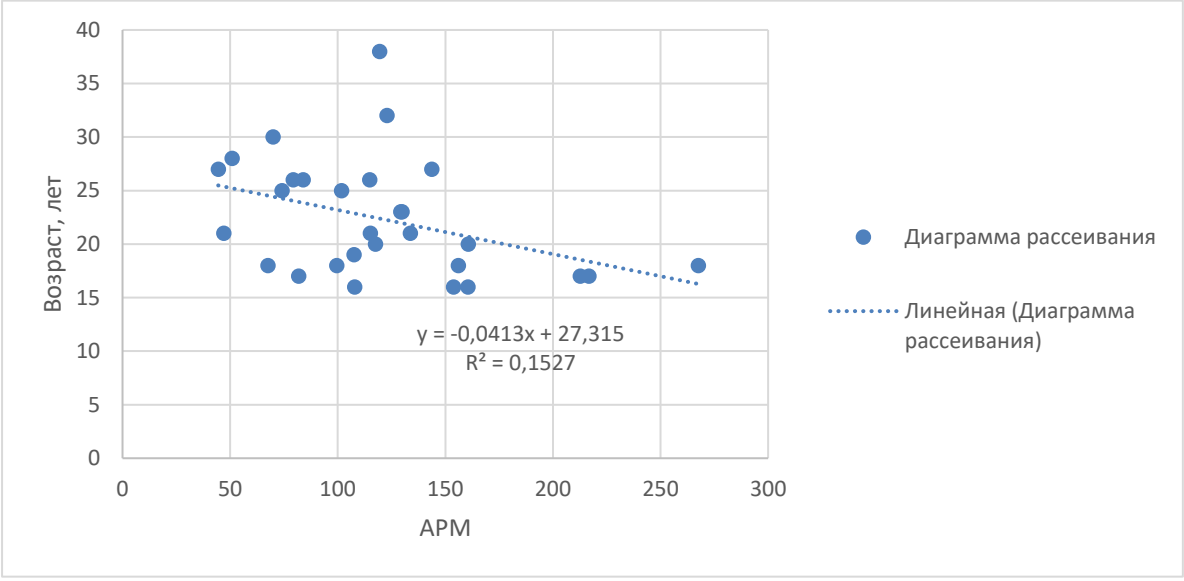

Рис.1. Линия тренда (линейная зависимость)

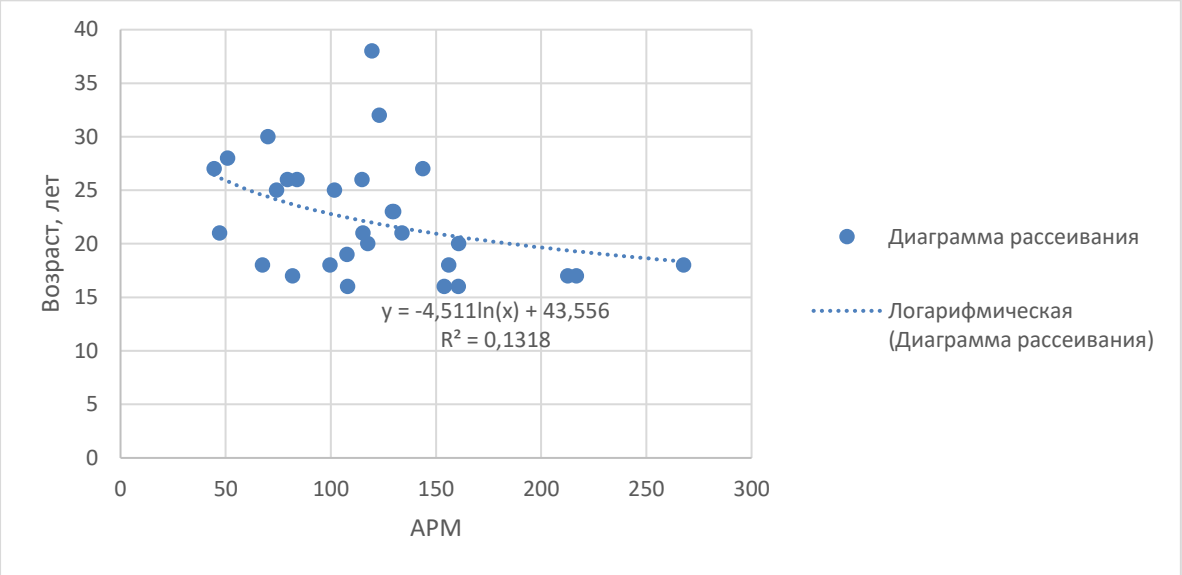

Рис.2. Линия тренда (логарифмическая зависимость)

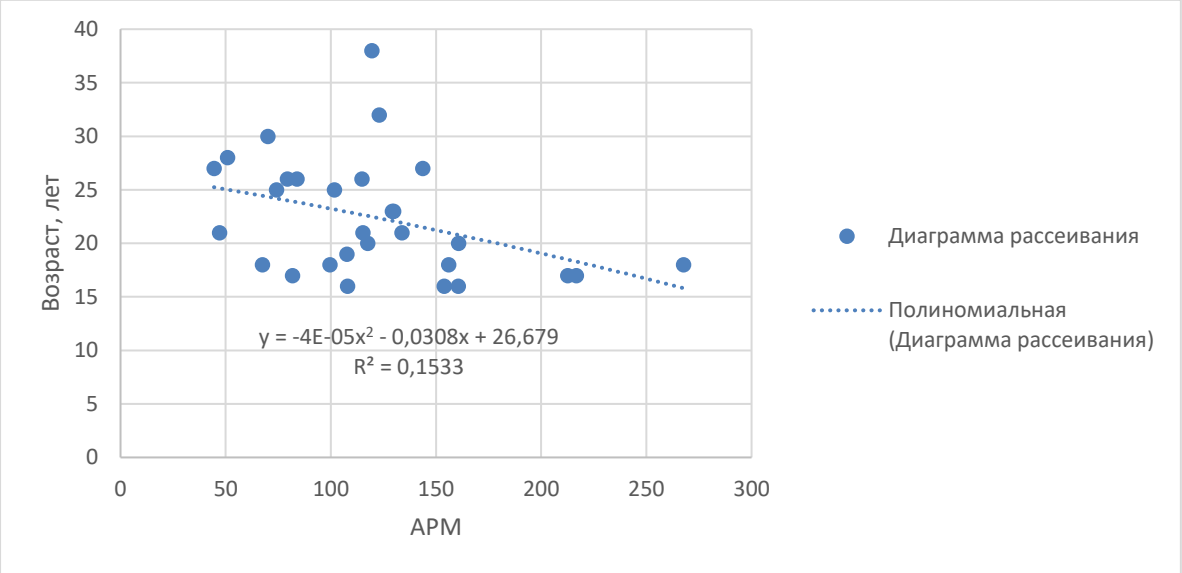

Рис.3. Линия тренда (квадратичная зависимость)

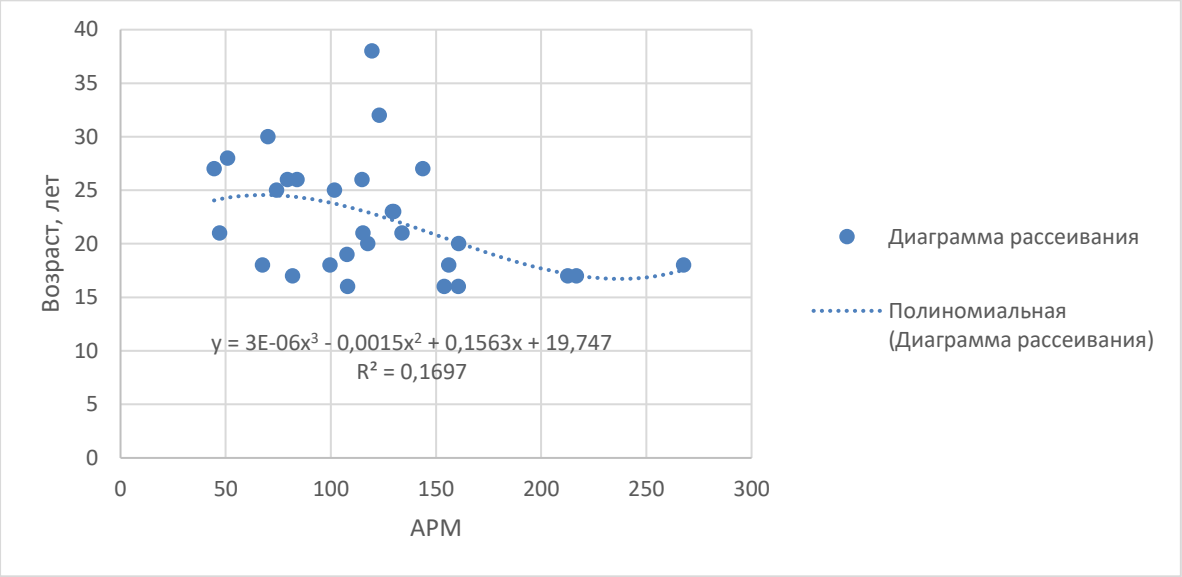

Рис.4. Линия тренда (кубическая зависимость)

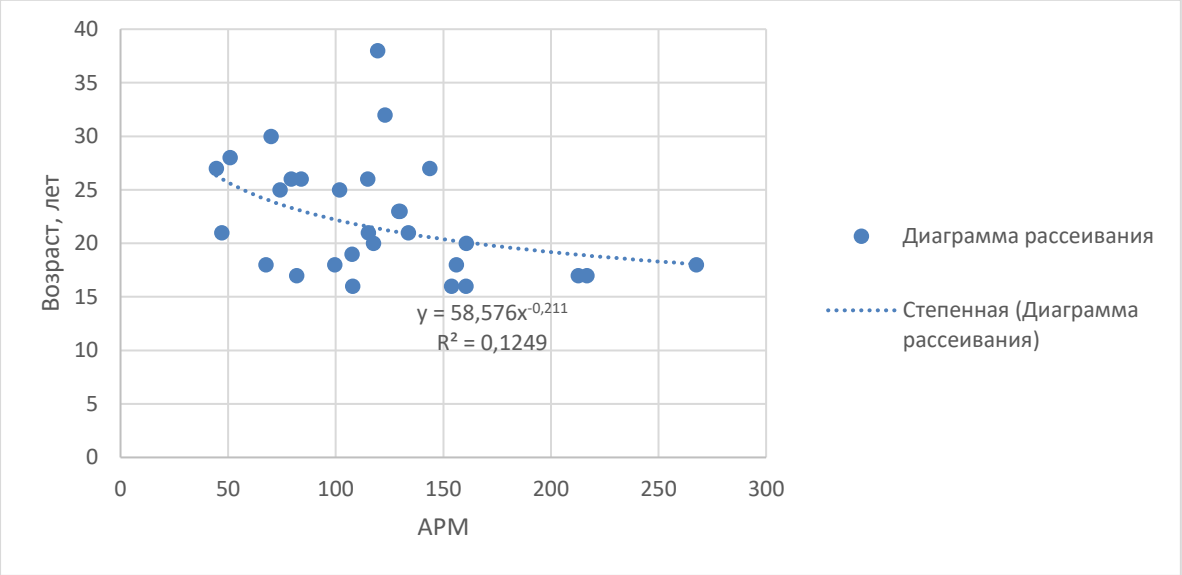

Рис.5. Линия тренда (степенная зависимость)

Обобщим полученную из графиков информацию в одну таблицу, которая будет содержать значения коэффициента корреляции R^2, среднего квадрата ошибки (СКО) и уравнение функции. Все значения были рассчитаны в MS Excel с

помощью формул СТАНДОТКЛОН для СКО или построение графика для R^2 и уравнения функции

Табл.2. Общая информация о разных видах зависимостей

y	R^2	Функция	СКО
x	0,7381	0,1521x	5,46
ax+b	0,1527	-0,0413x+27,315	2,14
ln(x)	0,1318	-4,511ln(x)+43,556	1,95
x^2	0,1533	-4E*0,5x^2 - 0,0308x + 26,679	0,86
x^3	0,1697	3E*6*x^3 - 0,0015x^2 + 0,1563x + 19,747	8,87
x^a	0,1249	58,576x^(-0,211)	2,05

Как мы видим, минимальный СКО у квадратичной функции -4E*05x^2 - 0,0308x + 26,679, а максимальное значение R^2 у линейной функции 0,1521x.

Теперь построим график остатков. Остаток - это разница между наблюдаемыми значениями и значениями, предсказанным моделью регрессии, а стандартизированный остаток найдем по формуле (1)

$$\text{Стандартизованный остаток} = \frac{\text{остаток}}{\sqrt{\frac{\text{остаток}^2}{n-1}}} \#1$$

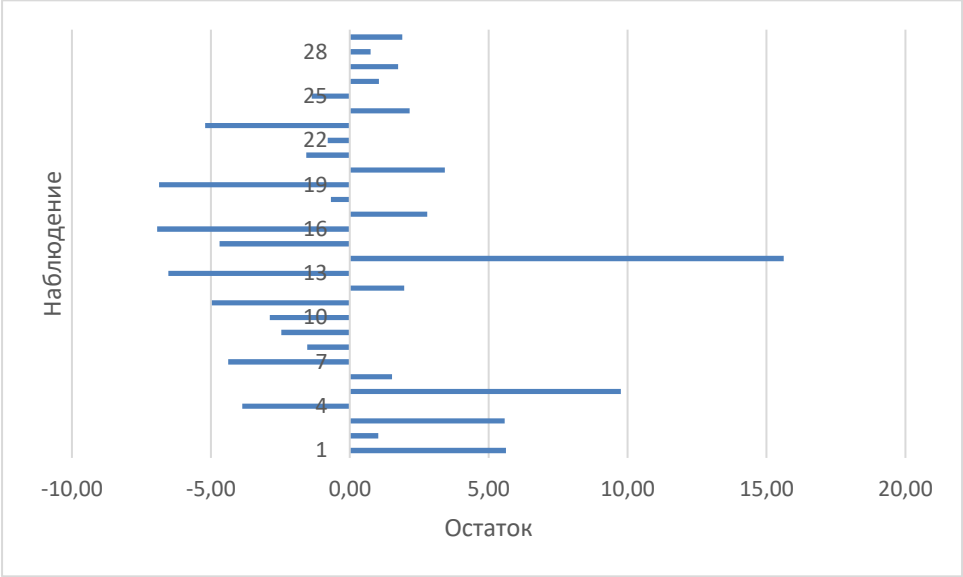

Рис.6. График остатков.

После построения графиков остатков рассчитаем критерий Дарбина-Уотсона - статистический критерий, используемый для тестирования автокорреляции первого порядка элементов исследуемой последовательности. Это можно сделать по формуле (2)

$$d = \frac{\sum_{t=2}^{n} \left(e_t - e_{t-1}\right)^2}{\sum_{t=1}^{n} e_t^2},$$

где et = yt - y’t, a y’t = ax+b

Расчеты показали, что d = 0,00053243, то есть d стремится к 0, а значит, что автокорреляция положительная.

Расстояние Кука оценивает эффект от удаления одного (рассматриваемого) наблюдения. Воспользуемся функцией в R для расчета расстояния Кука.

```
> cooksD_mdl <- lm(y~x1, df)
> cooksD <- cooks.distance(cooksD_mdl)
> cooksD
           1            2            3            4            5            6            7
0.0179687611 0.0757090194 0.0167806170 0.0058600827 0.0354440750 0.0476210528 0.0497449190
           8            9           10           11           12           13           14
0.0909956935 0.0009867820 0.0062478120 0.0036328785 0.0089685644 0.0572250977 0.2631934457
          15           16           17           18           19           20           21
0.0081734173 0.0484390742 0.0418215455 0.0100699634 0.0204825705 0.0003323371 0.0006670465
          22           23           24           25           26           27           28
0.0006753163 0.0152781283 0.0067998334 0.1008962641 0.0009876456 0.2336051293 0.0127113547
          29
0.0043019892
```

Рис.7. Расстояния Кука

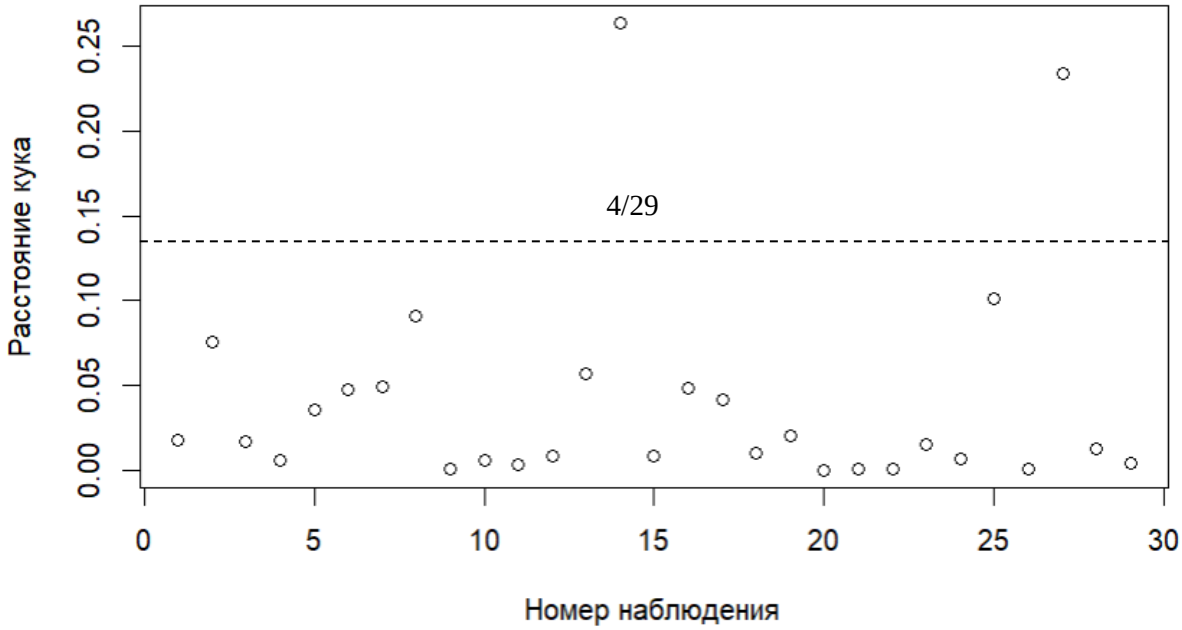

Рис.8. График расстояний Кука

Выбросы можно удалять по двум критериям:

- если расстояние Кука > 1, то это выброс
- если расстояние Кука > 4/n, то это выброс, где n - количество наблюдений

Так как 4/n = 4/29 = 0.1379, то 2 наблюдения будет выбросом. Пересчитаем расстояния Кука, выкинув 2 наблюдения и построим новый график расстояний.

```
> plot(cooksD, xlab="Номер наблюдения", ylab="Расстояние кука")
> cooksD_v2 <- cooksD[!(cooksD > 4/29)]
> cooksD_v2
           1            2            3            4            5            6            7
0.0179687611 0.0757090194 0.0167806170 0.0058600827 0.0354440750 0.0476210528 0.0497449190
           8            9           10           11           12           13           15
0.0909956935 0.0009867820 0.0062478120 0.0036328785 0.0089685644 0.0572250977 0.0081734173
          16           17           18           19           20           21           22
0.0484390742 0.0418215455 0.0100699634 0.0204825705 0.0003323371 0.0006670465 0.0006753163
          23           24           25           26           28           29
0.0152781283 0.0067998334 0.1008962641 0.0009876456 0.0127113547 0.0043019892
```

Рис.9. Расстояния Кука после удаления выброса (2 наблюдения)

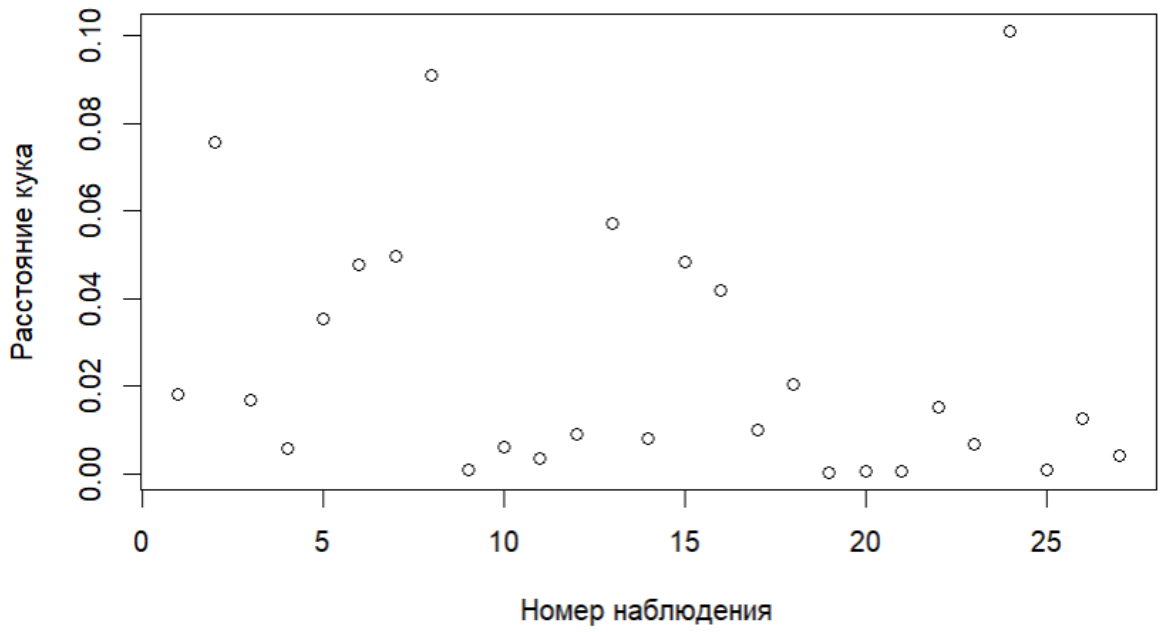

Рис.9. График расстояний Кука после удаления 2 наблюдений

Теперь критерий для нахождения выбросов равен 4/28 = 0.1429. Для всех наблюдений расстояние Кука < 0.1379, значит больше выбросов нет.

Расстояние Махаланобиса — это расстояние между двумя точками в многомерном пространстве. Он часто используется для поиска выбросов в статистическом анализе, включающем несколько переменных. Для расчета расстояния Махаланобиса воспользуемся встроенной в R функцией mahalanobis(df, colMeans(df), cov(df)), аргументы которой df - датасет, colMeans() - функция для подсчета среднего квадрата распределения, cov() - функция для построения ковариационной матрицы распределения.

```
> df = data.frame(y = c(143.43, 129.42, 69.42, 107.31, 122.21, 44.31, 46.42, 212.41, 117.11, 155.5
2, 153.51, 79.65, 67.43, 119.41, 160.34, 81.54, 50.42, 160.42, 107.41, 114.43, 115.31, 133.41, 99.4
3, 83.42, 216.42, 129.52, 267.42, 74.41, 101.42), x1 = c(27, 33, 30, 19, 32, 27, 21, 17, 20, 18, 1
6, 26, 18, 38, 16, 17, 28, 20, 16, 26, 21, 21, 18, 26, 17, 23, 18, 25, 21))
> head(df)
       y x1
1 143.43 27
2 129.42 33
3  69.42 30
4 107.31 19
5 122.21 32
6  44.31 27
> mahalanobis(df,colMeans(df),cov(df))
 [1] 1.18182811 3.97984142 1.94076518 0.61071298 3.06594059 2.17755591 2.64937408 3.35481599
 [9] 0.24380941 0.82752175 1.36711579 0.70681300 2.48395907 8.02469967 1.46605272 2.21916678
[17] 2.02546420 0.66004901 1.73426533 0.35939227 0.10922969 0.10965771 1.12094955 0.61956864
[25] 3.61901873 0.06015191 8.21518747 0.76972808 0.29736497
> df$mahal <- mahalanobis(df,colMeans(df),cov(df))
> df$p <- pchisq(df$mahal,df=28,lower.tail=FALSE)
```

Рис.10. Расчет расстояния Махаланобиса

Как видно, некоторые расстояния Махаланобиса намного больше других. Чтобы определить, является ли какое-либо из расстояний статистически значимым, нам нужно рассчитать их p-значения.

Значение p для каждого расстояния рассчитывается как значение p, которое соответствует статистике хи-квадрата расстояния Махаланобиса с k-1 степенями свободы, где k - количество переменных.

Табл.3. Расстояния Махаланобиса

	y	x1	mahal	p
1	143.43	27	1.18182811	0.75736590
2	129.42	33	3.97984142	0.26364914
3	69.42	30	1.94076518	0.58479167
4	107.31	19	0.61071298	0.89397556
5	122.21	32	3.06594059	0.38156975
6	44.31	27	2.17755591	0.53638268
7	46.42	21	2.64937408	0.44889931
8	212.41	17	3.35481599	0.34008586
9	117.11	20	0.24380941	0.97022505
10	155.52	18	0.82752175	0.84287350
11	153.51	16	1.36711579	0.71326075
12	79.65	26	0.70681300	0.87160031
13	67.43	18	2.48395907	0.47819701
14	119.41	38	8.02469967	0.04550399
15	160.34	16	1.46605272	0.69012721
16	81.54	17	2.21916678	0.52818299
17	50.42	28	2.02546420	0.56713837
18	160.42	20	0.66004901	0.88255811
19	107.41	16	1.73426533	0.62934289
20	114.43	26	0.35939227	0.94849723
21	115.31	21	0.10922969	0.99070728
22	133.41	21	0.10965771	0.99065380
23	99.43	18	1.12094955	0.77201934
24	83.42	26	0.61956864	0.89193833
25	216.42	17	3.61901873	0.30565068
26	129.52	23	0.06015191	0.99614638
27	267.42	18	8.21518747	0.04176760
28	74.41	25	0.76972808	0.85669312
29	101.42	21	0.29736497	0.96052329

Выбросом считается p-значение < 0.001. Как мы видим, в данной выборке выбросов нет.

Сводные данные по измененной модели после удаления выбросов

приведены в таблице 3.

Таблица 3. Сводные данные по лучшей модели парной регрессии

Характеристика	Значение
Уравнение	3E*6*x^3 - 0,0015x^2 + 0,1563x + 19,747
R^2	0,1697
СКО	8,87

Переходим ко второй части лабораторной работы.

Часть 2. Множественная линейная регрессия

Для этой части лабораторной работы мы будем анализировать уже 4 критерия, а не 1, как в первой части лабораторной работы.

Табл.4. Исходные данные

	y	x1	x2	x3	x4
Номер	АРМ	Возраст (лет)	Часов в неделю	Всего часов	Оценка по лиге
1	143,718	27	10	3000	5
2	129,232	23	10	5000	5
3	69,9612	30	10	200	4
4	107,602	19	20	400	3
5	122,891	32	10	500	3
6	44,457	27	6	70	2
7	46,9962	21	8	240	1
8	212,602	17	42	10000	7
9	117,488	20	14	2708	4
10	155,986	18	24	800	4
11	153,801	16	16	6000	3
12	79,2948	26	4	190	4

13	67,4754	18	12	350	3
14	119,437	38	6	1000	3
15	160,475	16	30	5000	5
16	81,7722	17	16	1500	5
17	50,8374	28	8	2000	4
18	160,646	20	10	120	5
19	107,912	16	14	350	5
20	114,781	26	28	1100	4
21	115,127	21	10	800	5
22	133,702	21	6	500	6
23	99,5088	18	20	800	5
24	83,9172	26	10	500	5
25	216,694	17	14	500	4
26	129,86	23	20	800	4
27	267,559	18	70	2520	6
28	74,1174	25	6	800	5
29	101,68	25	20	700	5

Для расчета матрицы парных коэффициентов корреляции загрузим датасет в датафрейм R и воспользуемся функцией cor().

Табл.5. Матрица парных коэффициентов корреляции

	y	x1	x2	x3	x4
y	1.0000000	-0.3477557	0.7137621	0.4514443	0.6273708
x1	-0.3477557	1.0000000	-0.4296975	-0.1889053	-0.2578981
x2	0.7137621	-0.4296975	1.0000000	0.3098969	0.4668981
x3	0.4514443	-0.1889053	0.3098969	1.0000000	0.4022994
x4	0.6273708	-0.2578981	0.4668981	0.4022994	1.0000000

Смотрим на уровень корреляции между коэффициентами выше 0.7, y и x2 зависимы.

Составим модель регрессии в R. В результате чего мы получаем уравнение регрессии y = -0.149*x1+1.960*x2+0.003*x3+13.19*x4 +26.93

```
> model <- lm(y~x1+x2+x3+x4, df)
> summary(model)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-48.347 -24.327  -4.767  24.112  75.974

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.935626  39.785293   0.677  0.50486
x1          -0.148920   1.211340  -0.123  0.90318
x2           1.960343   0.578222   3.390  0.00241 **
x3           0.003760   0.003134   1.200  0.24195
x4          13.199423   5.965632   2.213  0.03669 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.47 on 24 degrees of freedom
Multiple R-squared:  0.642,     Adjusted R-squared:  0.5824
F-statistic: 10.76 on 4 and 24 DF,  p-value: 3.856e-05
```

Рис.11. Построение модели регрессии

Для расчета значимости критериев модели необходимо использовать критерий Стьюдента. На рис.11. в столбце Pr(>|t|) содержится информация о значимой вероятности (p-value). Если p-значение меньше определенного уровня значимости (например, α = 0,005 или 0,01), то говорят, что переменная предиктора имеет статистически значимую связь с переменной ответа в модели.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.935626  39.785293   0.677  0.50486
x1          -0.148920   1.211340  -0.123  0.90318
x2           1.960343   0.578222   3.390  0.00241
x3           0.003760   0.003134   1.200  0.24195
x4          13.199423   5.965632   2.213  0.03669
```

Рис.12. Значения критерия Стьюдента

Найдем табличное значение критерия Стьюдента для 29-4-1=24 степеней свободы и доверительной вероятности 0.8 равно 1.3178. Так как расчетные значения для всех переменных, кроме x2 и x4, меньше табличного, то только переменная x2 и x4 является значимой.

F-статистика (критерий Фишера) - это критерий, используемая для проверки значимости коэффициентов регрессии в моделях линейной регрессии. f-статистика может быть рассчитана как MSR/MSE, где MSR представляет среднюю сумму регрессии квадратов, а MSE - средняя сумма ошибки квадратов.

Для данной модели расчетное значение критерия Фишера равен 10.76. Табличное значение найдем по формуле MS Excel =FРАСПОБР(0,001;1;29-2), принимая во внимание, что уровень значимости $\alpha = 0{,}001$. Табличное значение равно 13.61, что меньше расчетного значения, а значит модель является значимой.

Часть 3. Множественная нелинейная регрессия

Для множественной нелинейной регрессии рассмотрим различные виды регрессии: квадратичную, логарифмическую, кубическую. Переменная x2 остается в линейной зависимости в модели, так как мы выше посчитали, что по критерию Стьюдента это значимая переменная.

Построим квадратичную модель.

```
> model <- lm(formula = y ~ I(x1^2) + x2 + I(x3^2) + x4, data = df)
> summary(model)

Call:
lm(formula = y ~ I(x1^2) + x2 + I(x3^2) + x4, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-47.002 -22.500  -4.471  26.460  79.446

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.046e+01  3.010e+01   0.680  0.50311
I(x1^2)     3.414e-03  2.409e-02   0.142  0.88848
x2          2.014e+00  5.881e-01   3.425  0.00222 **
I(x3^2)     2.912e-07  3.589e-07   0.811  0.42525
x4          1.425e+01  5.985e+00   2.381  0.02552 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 34 on 24 degrees of freedom
Multiple R-squared:  0.6305,    Adjusted R-squared:  0.5689
F-statistic: 10.24 on 4 and 24 DF,  p-value: 5.556e-05
```

Рис.13. Квадратичная модель

Расчетный критерий 10.24, что меньше табличного значения (13.61), а значит модель является значимой.

Построим кубическую модель.

```
> model <- lm(formula = y ~ I(x1^3) + x2 + I(x3^3) + x4, data = df)
> summary(model)

Call:
lm(formula = y ~ I(x1^3) + x2 + I(x3^3) + x4, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-47.551 -19.804  -3.308  26.518  79.998

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.567e+01  2.789e+01   0.562  0.57945
I(x1^3)     2.482e-04  5.969e-04   0.416  0.68125
x2          2.097e+00  5.883e-01   3.563  0.00157 **
I(x3^3)     1.714e-11  3.820e-11   0.449  0.65757
x4          1.499e+01  6.029e+00   2.485  0.02030 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 34.21 on 24 degrees of freedom
Multiple R-squared:  0.626,     Adjusted R-squared:  0.5636
F-statistic: 10.04 on 4 and 24 DF,  p-value: 6.383e-05
```

Рис.14. Кубическая модель

Критерий Стьюдента для каждой переменной рассчитан в столбце Pr(>|t|). Расчетное значение критерия Фишера равно 10.04, модель является значимой.

Построим логарифмическую модель.

```
> model <- lm(formula = y ~ log(x1) +x2 + log(x3) + x4, data = df)
> summary(model)

Call:
lm(formula = y ~ log(x1) + x2 + log(x3) + x4, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-51.941 -21.744  -4.167  23.158  71.024

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  29.5894   103.2425   0.287  0.77688
log(x1)     -12.4244    29.2644  -0.425  0.67494
x2            1.9457     0.5827   3.339  0.00274 **
log(x3)       6.0789     6.0384   1.007  0.32412
x4           12.7010     6.2591   2.029  0.05367 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.63 on 24 degrees of freedom
Multiple R-squared:  0.6385,    Adjusted R-squared:  0.5782
F-statistic:  10.6 on 4 and 24 DF,  p-value: 4.317e-05
```

Рис.15. Логарифмическая модель

Табличное значение критерия Фишера равно 10.6, а расчетное - 13.61, что меньше табличного, следовательно модель является значимой.

Построим модель степенной зависимости.

```
> model <- lm(formula = log(y)~ log(x1) + x2 + log(x3) + x4, data = df)
> summary(model)

Call:
lm(formula = log(y) ~ log(x1) + x2 + log(x3) + x4, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.65970 -0.21446  0.00265  0.19693  0.50162

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.858106   0.954449   4.042 0.000474 ***
log(x1)     -0.134016   0.270541  -0.495 0.624850
x2           0.011168   0.005387   2.073 0.049049 *
log(x3)      0.072374   0.055824   1.296 0.207136
x4           0.132274   0.057864   2.286 0.031383 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3109 on 24 degrees of freedom
Multiple R-squared:  0.5744,    Adjusted R-squared:  0.5035
F-statistic: 8.098 on 4 and 24 DF,  p-value: 0.0002788
```

Рис.16. Экспоненциальная модель

Табличное значение критерия Фишера равно 13.61, а расчетное – 8.098, что меньше табличного, следовательно модель является значимой.

Составим сводную таблицу по всем регрессионным моделям.

Табл.6. Регрессионные модели

Модель	Уравнение регрессии	R^2	СКО	Критерий Фишера
Множественная линейная	-0.14*x1+1.96*x2+0.003*x3-13.19*x4+26	0.64	33.4	10.76
Множественная квадратичная	20+ 0.003*x1^2 + 2.01*x2 + 14.25 *x3^2 + 14.2*x4	0.63	34	10.24
Множественная кубическая	1.5*x1^3 + 2.08*x2 + 0.000000014*x3^3 + 1.499 *x4 +1.57	0.62	34.2	10.04
Множественная логарифмическая	-12.42*log(x1) + 1.94*x2 +6.079*log(x3) – 12.7*x4 + 29.043	0.63	33.6	10.64
Множественная степенная	x1^(-0.13) + 0.01*x2 + x3^(0.07) + 0.13*x4+ 3.85	0.57	0.31	8.09

Заключение

Изучен расчет основных статистик (среднее значение (мат. ожидание), стандартное отклонение, дисперсию, минимум, максимум, моду, медиану, асимметрию, эксцесс, размах, квартили) и построение основных видов графиков (график зависимости от параметра, гистограмма частот, диаграмма размахов, кривые распределений) по случайной выборке.

Также были выполнены все регрессионного анализа, в том числе линейная и нелинейная, множественная линейная и множественная нелинейная регрессия. . В ходе работы было выяснено, что количество часов в неделю влияет на средний АРМ игрока.