

1A	1B	1C	1D	1E	1F	2A	2B	2C	2D	2E

1 Задание 1

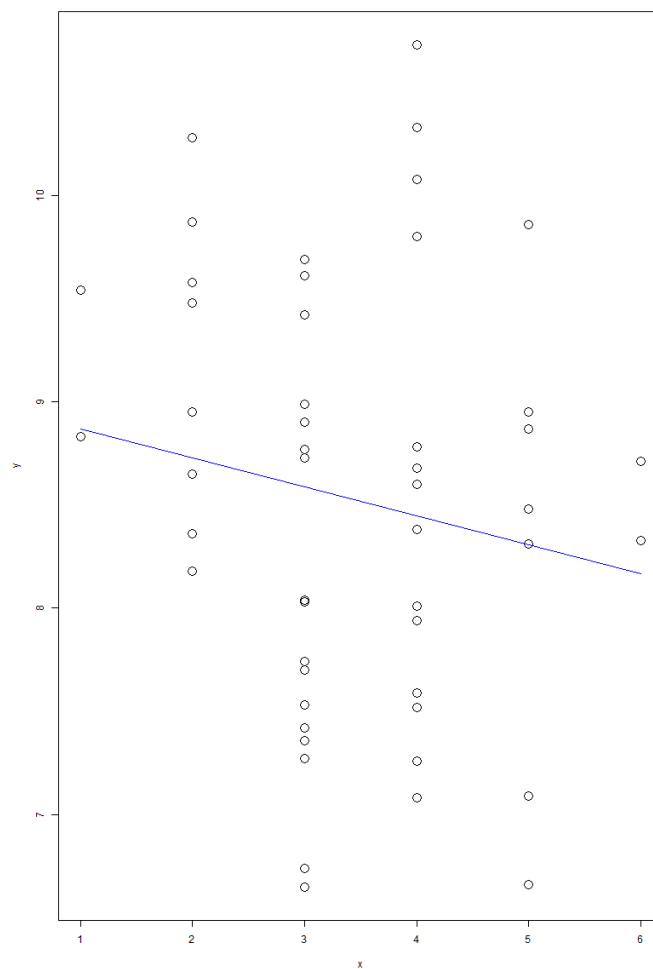
1. Результаты статистического эксперимента приведены в таблице 1. Требуется оценить характер зависимости наблюдаемой переменной Y от ковариаты X .
 - a) Построить графически результаты эксперимента. Сформулировать линейную регрессионную модель переменной Y по переменной X . Построить МНК оценки параметров сдвига β_1 и масштаба β_2 . Построить полученную линию регрессии. Оценить визуально соответствие полученных данных и построенной оценки.
 - b) Сформулировать полиномиальную модель, включающую дополнительный член с X^2 . Построить МНК оценки параметров $\beta_1, \beta_2, \beta_3$ в данной модели. Изобразить графически полученную регрессионную зависимость. Оценить визуально соответствие полученных данных и построенной оценки.
 - c) На базе ошибок полиномиальной модели построить гистограмму. Проверить значимость отклонения от нормального распределения по χ^2 . Визуально оценить данный факт.
 - d) В предположении нормальности построить частные и совместные доверительные интервалы для параметров β_2 и β_3 уровня доверия $1 - \alpha$.
 - e) Сформулировать гипотезы линейности зависимости и независимости наблюдаемой переменной Y от ковариаты X . Провести проверку значимости.
 - f) С использованием AIC и BIC выбрать наилучшую модель.
 - g) Интерпретировать полученные результаты. Написать отчет.

Таблица 1 $\alpha = 0.20; h = 0.94$.

No	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Y	7.70	8.65	8.87	8.90	7.74	8.01	9.86	8.60	9.87	7.08	8.31	10.28	9.54	9.48	10.08	8.95	8.38
X	3	2	5	3	3	4	5	4	2	4	5	2	1	2	4	2	4
No	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
Y	7.59	9.42	8.77	8.36	9.80	9.69	8.99	7.52	8.73	9.58	8.95	6.74	8.18	7.53	7.36	7.26	6.66
X	4	3	3	2	4	3	3	4	3	2	5	3	2	3	3	4	5
No	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	
Y	8.04	8.03	7.42	8.48	8.33	8.71	10.73	6.65	7.27	8.78	10.33	8.68	7.94	8.83	9.61	7.09	
X	3	3	3	5	6	6	4	3	3	4	4	4	4	1	3	5	

1.1 Задание А

На рис. 1.1 изображен график результатов линейной модели с исходными данными.

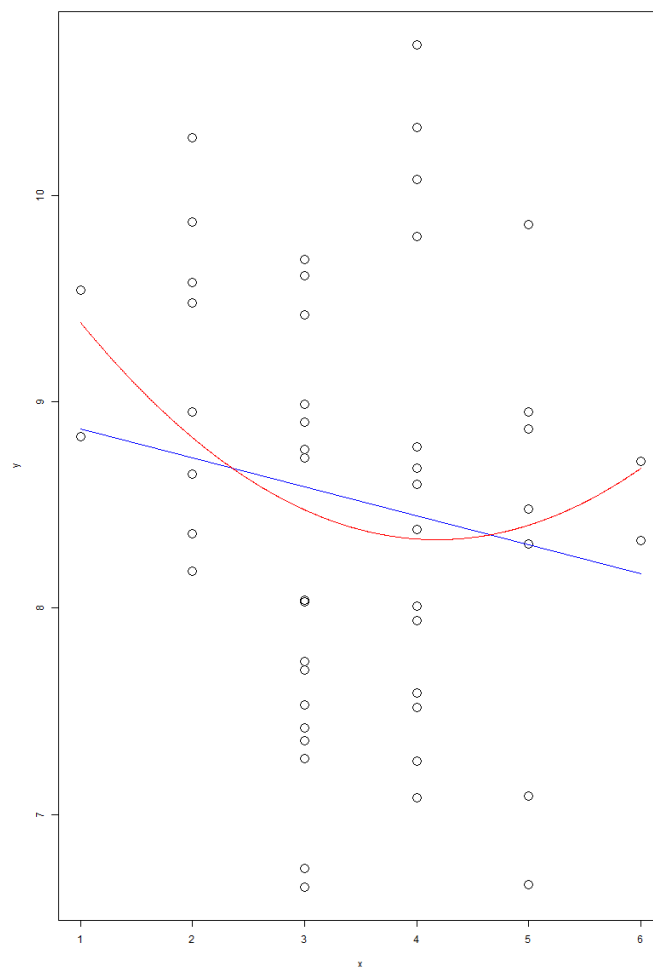


Полученная оценка:

$$\hat{\beta}_1 = 9.0116707, \quad \hat{\beta}_2 = -0.1408926.$$

1.2 Задание В

На рис. 1.2 изображен **полиномиальная модель** в сравнении с **линейной моделью**.



Полученные оценки полиномиальной модели:

$$\hat{\beta}_1 = 10.1505353, \quad \hat{\beta}_2 = -0.8704798, \quad \hat{\beta}_3 = 0.1041729.$$

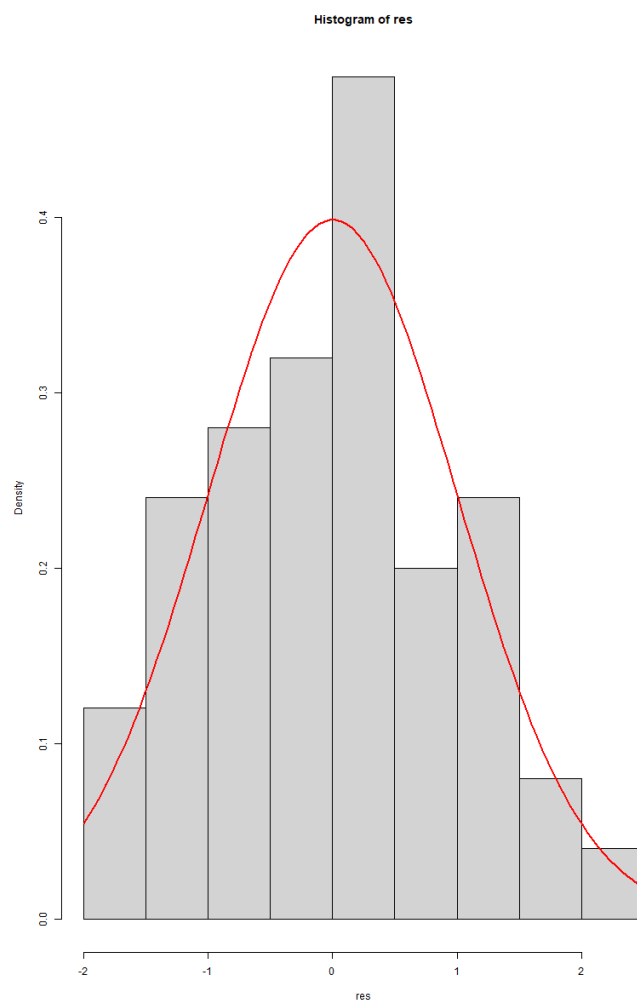
1.3 Задание С

Упорядоченные остатки полиномиальной модели:

1	2	3	4	5	6
-0.77665193	-0.17626731	0.46754157	0.42334807	-0.73665193	-0.32538230
7	8	9	10	11	12
1.45754157	0.26461770	1.04373269	-1.25538230	-0.09245843	1.45373269
13	14	15	16	17	18
0.15577156	0.65373269	1.74461770	0.12373269	0.04461770	-0.74538230
19	20	21	22	23	24
0.94334807	0.29334807	-0.46626731	1.46461770	1.21334807	0.51334807
25	26	27	28	29	30
-0.81538230	0.25334807	0.75373269	0.54754157	-1.73665193	-0.64626731

31	32	33	34	35	36
-0.94665193	-1.11665193	-1.07538230	-1.74245843	-0.43665193	-0.44665193
37	38	39	40	41	42
-1.05665193	0.07754157	-0.34788032	0.03211968	2.39461770	-1.82665193
43	44	45	46	47	48
-1.20665193	0.44461770	1.99461770	0.34461770	-0.39538230	-0.55422844
49	50				
1.13334807	-1.31245843				

Для остатков полиномиальной модели была построена гистограмма с шагом $h = 1.5$, она изображена на рис. 1.3. Красной линией обозначен график плотности нормального распределения $N(0, s^2)$.



Ниже приведен фрагмент, показывающий информацию о приведенной выше гистограмме.

```
> h$breaks
[1] -2.0 -1.5 -1.0 -0.5  0.0  0.5  1.0  1.5  2.0  2.5
> h$counts
[1]  3  6  7  8 12  5  6  2  1
```

С помощью программной нелинейной оптимизации был вычислен минимум статистики X^2 . Полученное значение $X_{min}^2 = 0.3941346$.

В данном случае количество интервалов равно 6, гипотеза является сложной и имеет размерность параметра 1. При выполнении гипотезы нормальности статистика X^2 должно иметь распределение χ_{6-1-1}^2 .

Статистика хи-квадрат: -84.49288. Критическое значение (alpha=0.2): 5.988617. Не отвергаем H_0 : остатки нормальны (на уровне значимости 0.2).

1.4 Задание D

Доверительные интервалы:

$$-1.65844087 < \beta_2 < -0.08251865$$

$$-0.00589722 < \beta_3 < 0.21424297$$

1.5 Задание E

Model 1: $Y \sim X$

Model 2: $Y \sim I(X) + I(X^2)$

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
2	47	49.098	1	1.5809	1.5134	0.2248
21	47	49.098	2	2.8974	1.3868	0.2599

1. Проверка линейности зависимости

Гипотезы:

- H_0 : Модель $Y \sim X$ (линейная) адекватна
- H_1 : Модель $Y \sim X + X^2$ (квадратичная) лучше

Результаты:

- $F = 1.5134$
- Критическое значение $F_{1,47,0.8} \approx 1.6658$
- p -значение = 0.2248

Вывод: Поскольку $F = 1.5134 < F_{1,47,0.8} \approx 1.689545$ (и $p > 0.2$), **не отвергаем H_0** .
 \Rightarrow Нет статистически значимых оснований включать квадратичный член.

2. Проверка значимости ковариаты X

Гипотезы:

- H_0 : Модель без предиктора ($Y \sim 1$) адекватна
- H_1 : Модель с X лучше (линейная или квадратичная)

Результаты:

- $F = 1.3868$
- Критическое значение $F_{2,47,0.8} \approx 1.6658$
- p -значение = 0.2599

Вывод: Поскольку $F = 1.3868 < F_{2,47,0.8} \approx 1.66583$ (и $p > 0.2$), **не отвергаем H_0** .
 \Rightarrow Ковариата X не оказывает значимого влияния.

Общий вывод

1. Линейная модель $Y \sim X$ адекватна (квадратичный член не требуется)
2. Ковариата X не является статистически значимой
3. При $\alpha = 0.2$ ни одна из альтернативных гипотез не подтвердилась

1.6 Задание F

AIC для не зависящей от x модели: 147.8505

AIC для линейной модели: 148.5682

AIC для квадратичной модели: 148.9837

BIC для не зависящей от x модели: 151.6746

BIC для линейной модели: 154.3043

BIC для квадратичной модели: 156.6318

Лучшей моделью по критерию Акаике оказалась модель (3), в которой y не зависит от ковариаты. Аналогично по Байесовскому критерию.

2 Задание 2

task2.png

Рис. 1: Исходные данные задания 2

2.1 Задание А

В исходных данных есть два фактора A и B с количеством уровней $d_1 = 5$ и $d_2 = 2$ соответственно. В качестве взвешивания по обоим факторам возьмем первый базовый уровень (то есть веса нулевые для всех значений факторов, кроме первого). Модель двухфакторного анализа выглядит следующим образом:

$$E(Y|A = a_i, B = b_j) = \eta_{ij} = \mu + \alpha_i^{(1)} + \alpha_j^{(2)} + \alpha_{ij}^{(12)},$$

где $i = 1, 2, 3, 4, 5$, $j = 1, 2$, ограничения: $\alpha_1^{(1)} = 0$, $\alpha_1^{(2)} = 0$, $\alpha_{i1}^{(12)} = \alpha_{1j}^{(12)} = 0$ для любых

допустимых i и j и $\mu = \eta_{11}$.

Таким образом у модели 10 свободных параметров: μ , $\alpha_2^{(1)}$, $\alpha_3^{(1)}$, $\alpha_4^{(1)}$, $\alpha_5^{(1)}$, $\alpha_2^{(2)}$, $\alpha_{22}^{(12)}$, $\alpha_{32}^{(12)}$, $\alpha_{42}^{(12)}$, $\alpha_{52}^{(12)}$.

Найдем оценки параметров с помощью метода наименьших квадратов:

μ	$\alpha_2^{(1)}$	$\alpha_3^{(1)}$	$\alpha_4^{(1)}$	$\alpha_5^{(1)}$	$\alpha_2^{(2)}$	$\alpha_{22}^{(12)}$	$\alpha_{32}^{(12)}$	$\alpha_{42}^{(12)}$	$\alpha_{52}^{(12)}$
19.136	0.534	-2.838	-2.436	-2.032	-0.332	1.866	1.002	-0.058	0.482

В данной модели ранг регрессора $r = 10$, размер выборки $n = 50$, тогда оценка дисперсии вычисляется следующим образом: $s^2 = SSe/(n - r) = SSe/40$.

Вычисленные значения:

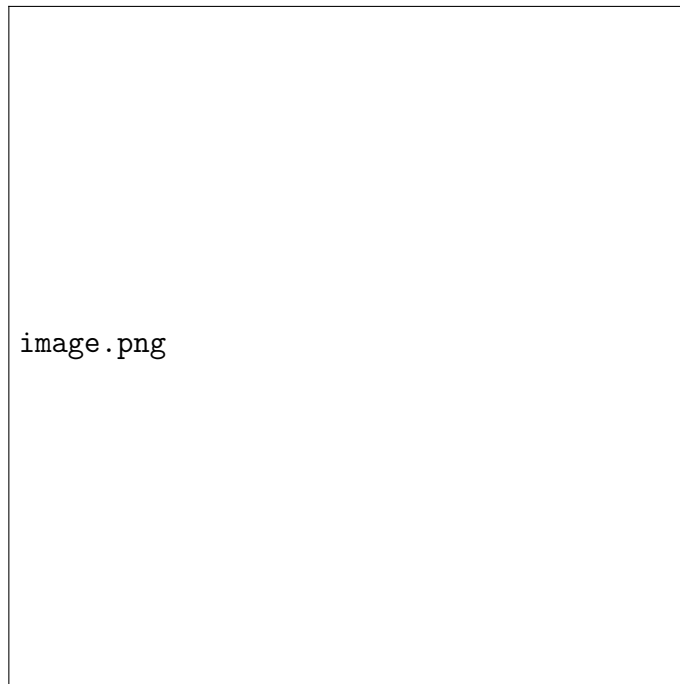
$$SSe = 14.43896, \quad s^2 = 0.360974.$$

2.2 Задание В

Предположение аддитивности $H_{(12)}$ формулируется следующим образом:

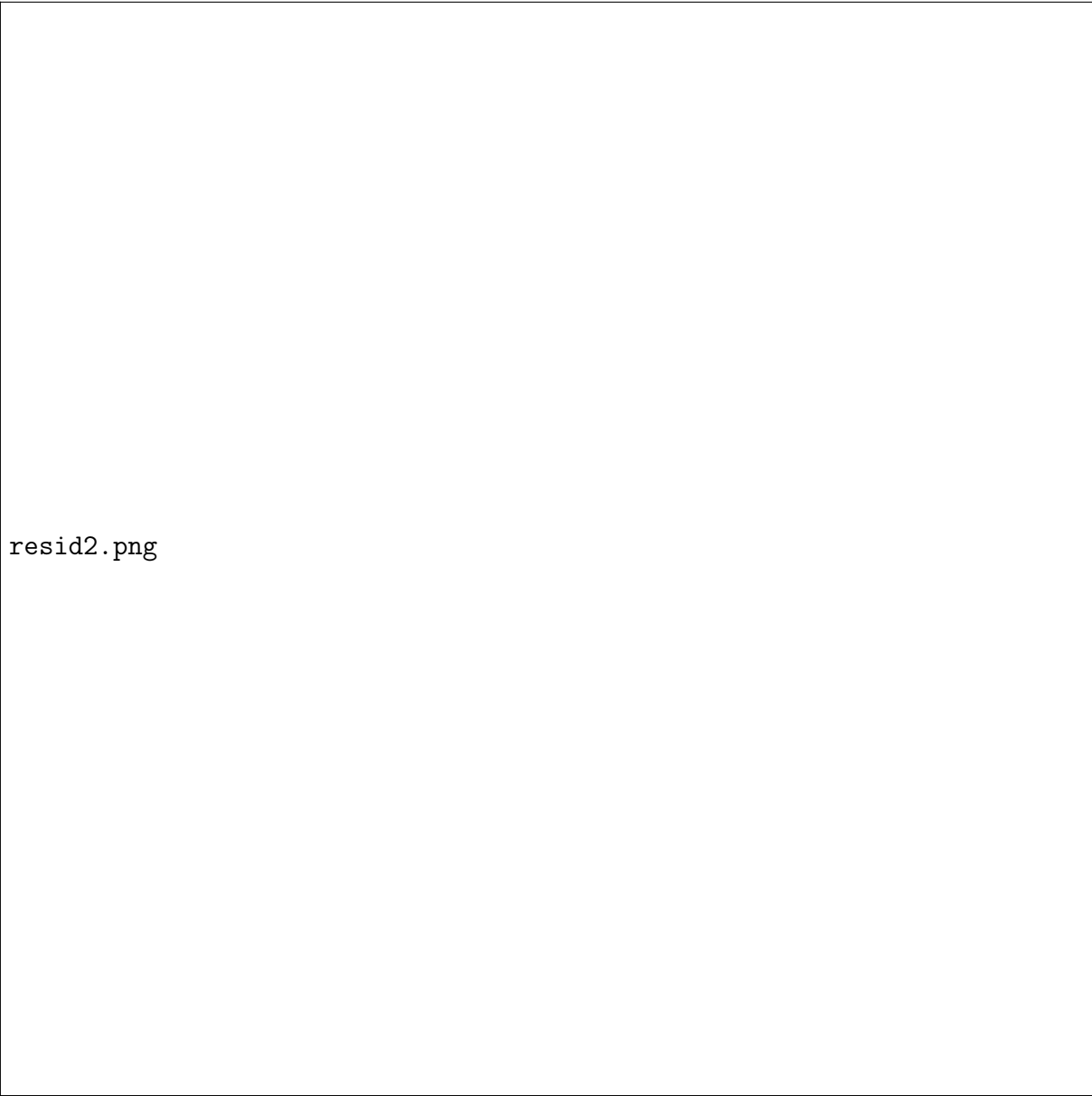
$$H_{12} : a_{ij}^{(12)} = 0 \text{ для всех } i, j. \quad (1)$$

Проверим визуально, согласуются ли данные с этой гипотезой. Для этого построим η_{ij} от i для каждого фиксированного уровня j . Если выполнена $H_{(12)}$, то все графики для различных уровней j должны получаться вертикальным сдвигом одного графика.



На рисунке видно, что графики схожи, однако не являются вертикальными сдвигами одного графика. Можно сделать вывод об отсутствии аддитивности факторов.

2.3 Задание С



resid2.png

Рис. 2: Значение ошибок модели

Для этих значений было проведена аналогичная проверка согласованности с нормальным распределением с помощью критерия χ^2 . Для этого остатки были разбиты на 10 групп по 5 штук, была выполнена минимизация статистики X^2 , оптимальное полученное значение $X_{min}^2 = 5.525553$.

При выполнении гипотезы статистика распределена по распределению χ^2 с $10 - 2 - 1 = 7$ степенями свободы, P -значение у гипотезы согласования $PV = 0.355152$. Таким образом отвергнуть гипотезу о нормальности можно только при уровне значимости больше 0.355152.

На рис. 3 изображена гистограмма с шагом $h = 0.61$. Красной линией изображена плотности нормального распределения $N(0, s^2)$.

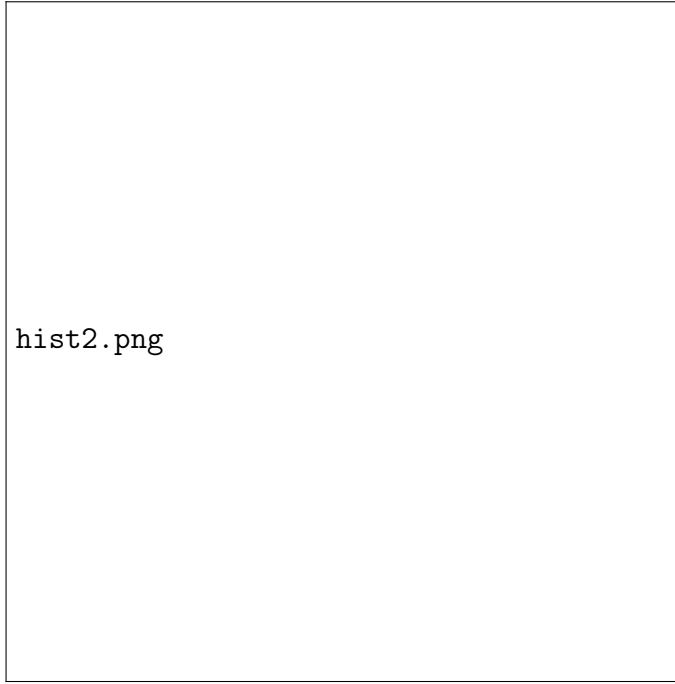


Рис. 3: Гистограмма ошибок

Можно наблюдать некоторое сходство гистограммы с нормальным распределением, однако нельзя говорить о полном сходстве с графиком.

2.4 Задание D

В рамках проведения двухфакторного дисперсионного анализа проверим выполнение следующих гипотез:

- Отсутствие взаимодействия факторов
 $H_{(12)} : \alpha_{ij}^{(12)} = 0$, где $i = 1, \dots, d_1, j = 1, \dots, d_2$;
- Отсутствие влияния фактора А на результат (при выполнении H_{12})
 $H_{(1)} : \alpha_i^{(1)} = \alpha_{ij}^{(12)} = 0$, где $i = 1, \dots, d_1$;
- Отсутствие влияния фактора В на результат (при выполнении H_{12})
 $H_{(2)} : \alpha_j^{(2)} = \alpha_{ij}^{(12)} = 0$, где $j = 1, \dots, d_2$;
- Отсутствие влияния факторов (при выполнении H_{12}, H_1, H_2)
 $H_{(0)} : \alpha_i^{(1)} = \alpha_j^{(2)} = \alpha_{ij}^{(12)} = 0$, где $i = 1, \dots, d_1, j = 1, \dots, d_2$.

Для проверки гипотез будем использовать F -статистику:

$$F = \frac{SS_H/q}{SS_e/(n-r)},$$

где SS_H — сумма квадратов ошибок для модели, полученной при выполнении гипотезы, q — размерности оценивания в гипотезе, SS_e — сумма квадратов ошибок в исходной модели, n — размер выборки, r — ранг матрицы регрессора.

Эта статистика при выполненной гипотезе имеет распределение Фишера-Снедекора со степенями свободы $q, n - r$.

Для гипотезы $H_{(12)}$ значение $q = (d_1 - 1)(d_2 - 1) = 4$, для $H_{(1)}$ $q = d_2(d_1 - 1) = 8$, для $H_{(2)}$ $q = d_1(d_2 - 1) = 5$, для $H_{(0)}$ $q = d_1d_2 - 1 = 9$.

Значение r для всех гипотез будет равно $d_1d_2 = 10$, таким образом $n - r = 40$. В задании А было вычислено значение $SS_e = 14.43896$.

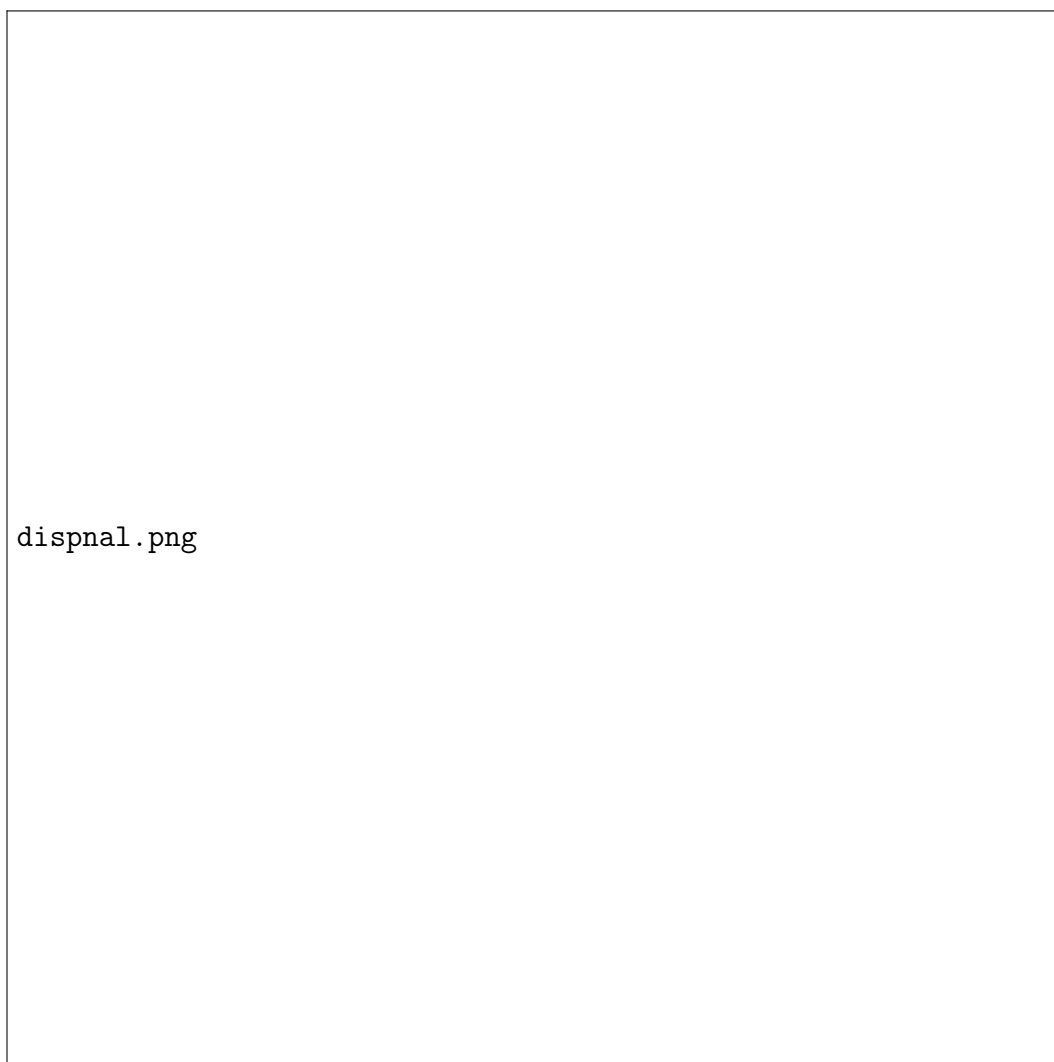


Рис. 4: Данные дисперсионного анализа

2.5 Задание Е

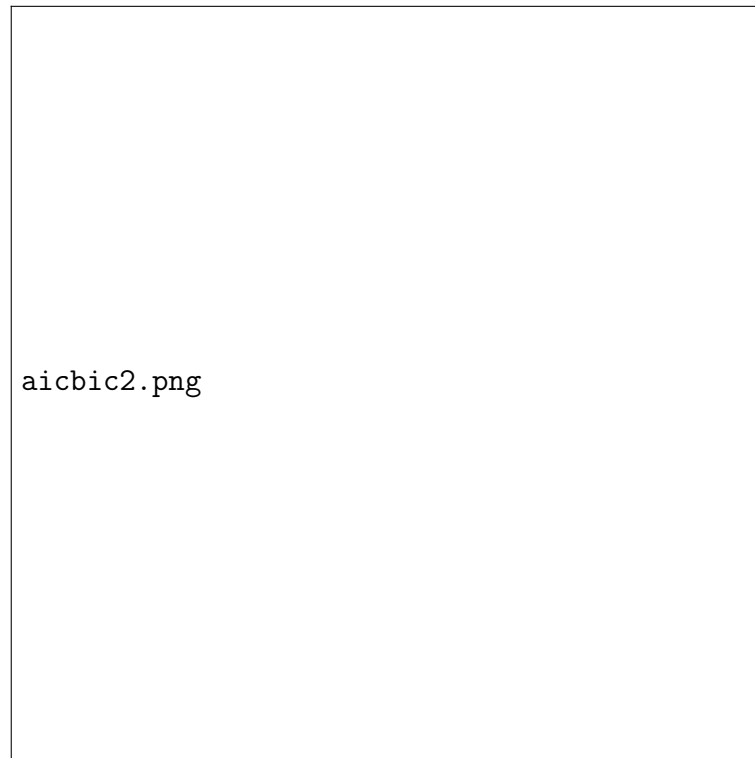


Рис. 5: Сравнение моделей $H_{(12)}$, $H_{(1)}$, $H_{(2)}$, $H_{(0)}$

По критерию Акаике лучшей является аддитивная модель, что можно было угадать из значений F -статистики. Однако по Байесовскому критерию лучшей является $H_{(2)}$, но аддитивная тоже близка по значению критерия.

Тем не менее исходная модель имеет значение $AIC = 101.7892$ и $BIC = 122.8215$, так что, несмотря на то, что она сложнее остальных моделей, она является предпочтительной и по информационным критериям.

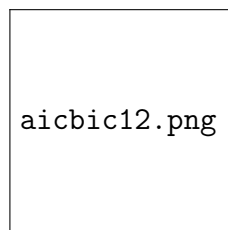


Рис. 6: Информационные критерии для исходной модели