Большая компьютерная работа

Выполнили:
Бахтиарова Кристина
Журавлева Кристина
Руденко Александра
Саргсян Нарек
Хромыщенко Иван

Постановка задачи

Описание и обоснование системы показателей

Источник данных: Kaggle.com

Данные: bodyfat.csv

Характеристика данных:

- 252 наблюдения (измерения тела мужчин)
- целевая переменная - ‘BodyFat’
- 13 непрерывных объясняющих переменных

Уравнение Сири:
100*BodyFat = 495/Density - 450

Таблица 1. Описание данных.

Переменная	Описание переменной, англ	Описание переменной, рус
Зависимая переменная		
BodyFat	Percent body fat from Siri's (1956) equation	Процент жира в организме по уравнению Сири (1956)
Объясняющие переменные		
Age	Age (years)	Возраст (в годах)
Weight	Weight (lbs)	Вес (фунты)
Height	Height (inches)	Рост (дюймы)
Neck	Neck circumference (cm)	Обхват шеи (см)
Chest	Chest circumference (cm)	Обхват грудной клетки (см)
Abdomen	Abdomen circumference (cm)	Обхват живота (см)
Hip	Hip circumference (cm)	Обхват бедер (см)
Thigh	Thigh circumference (cm)	Обхват ляжки (см)
Knee	Knee circumference (cm)	Обхват колена (см)
Ankle	Ankle circumference (cm)	Обхват лодыжки (см)
Biceps	Biceps (extended) circumference (cm)	Обхват бицепса (вытянутого) (см)
Forearm	Forearm circumference (cm)	Обхват предплечья (см)
Wrist	Wrist circumference (cm)	Обхват запястья (см)

Обоснование репрезентативности выборки

Таблица 2. Первые 10 наблюдений выборки.

BodyFat <dbl>	Age <int>	Weight <dbl>	Height <dbl>	Neck <dbl>	Chest <dbl>	Abdomen <dbl>	Hip <dbl>	Thigh <dbl>
12.3	23	154.25	67.75	36.2	93.1	85.2	94.5	59.0
6.1	22	173.25	72.25	38.5	93.6	83.0	98.7	58.7
25.3	22	154.00	66.25	34.0	95.8	87.9	99.2	59.6
10.4	26	184.75	72.25	37.4	101.8	86.4	101.2	60.1
28.7	24	184.25	71.25	34.4	97.3	100.0	101.9	63.2
20.9	24	210.25	74.75	39.0	104.5	94.4	107.8	66.0
19.2	26	181.00	69.75	36.4	105.1	90.7	100.3	58.4
12.4	25	176.00	72.50	37.8	99.6	88.5	97.1	60.0
4.1	25	191.00	74.00	38.1	100.9	82.5	99.9	62.9
11.7	23	198.25	73.50	42.1	99.6	88.6	104.1	63.1

Продолжение таблицы 2.

Neck <dbl>	Chest <dbl>	Abdomen <dbl>	Hip <dbl>	Thigh <dbl>	Knee <dbl>	Ankle <dbl>	Biceps <dbl>	Forearm <dbl>	Wrist <dbl>
36.2	93.1	85.2	94.5	59.0	37.3	21.9	32.0	27.4	17.1
38.5	93.6	83.0	98.7	58.7	37.3	23.4	30.5	28.9	18.2
34.0	95.8	87.9	99.2	59.6	38.9	24.0	28.8	25.2	16.6
37.4	101.8	86.4	101.2	60.1	37.3	22.8	32.4	29.4	18.2
34.4	97.3	100.0	101.9	63.2	42.2	24.0	32.2	27.7	17.7
39.0	104.5	94.4	107.8	66.0	42.0	25.6	35.7	30.6	18.8
36.4	105.1	90.7	100.3	58.4	38.3	22.9	31.9	27.8	17.7
37.8	99.6	88.5	97.1	60.0	39.4	23.2	30.5	29.0	18.8
38.1	100.9	82.5	99.9	62.9	38.3	23.8	35.9	31.1	18.2
42.1	99.6	88.6	104.1	63.1	41.7	25.0	35.6	30.0	19.2

Мужчины имеют:

1) Разный возраст
2) Разное телосложение

Гипотезы исследования

1) зависимость процента жира в организме (‘Bodyfat’) от различных измерений тела;
2) наличие аномальных наблюдений в выборке;
3) зависимая переменная ‘BodyFat’ имеет нормальное распределение.

Основные характеристики СВ и диагностика выбросов

Характеристики СВ и диагностика выбросов

Были получены следующие выводы:

- среднее, мода и медиана примерно равны – предполагаем наличие нормального распределения исследуемой переменной **BodyFat**;
- коэффициент вариации выше 33% означает, что переменная **BodyFat** неоднородна;
- по правилу 1,5 IQR был определен **один выброс** (47,5);
- по правилу 3 IQR был определен **выбросов не обнаружено**;
- по правилу трёх сигм был определен **один выброс** (47,5)**.**

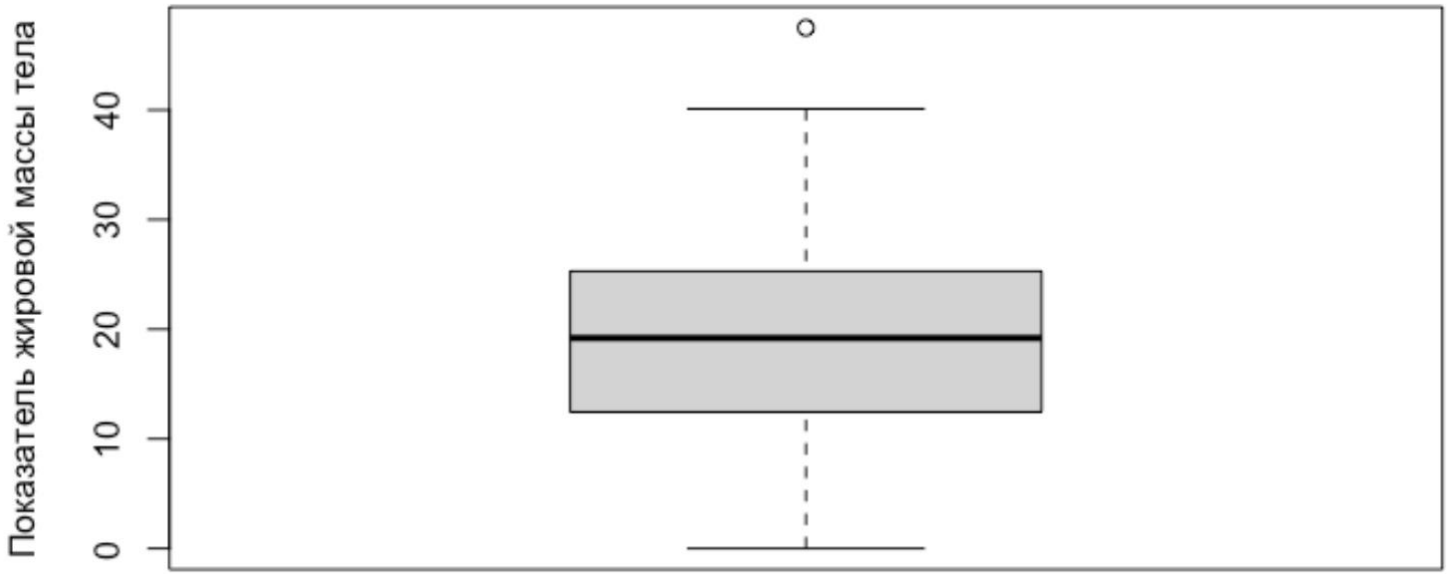


Рисунок 1. Boxplot.

Диагностика выбросов

Тест Граббса

Проверяли максимальное (47,5) и минимальное (0) значения.

Выбросами они **не являются**.

Тест Рознера

Проверяли значение, являющееся выбросом по правилу 1,5 IQR (47,5).

Выбросом оно **не является**.

Проверка соответствия эмпирического распределения нормальному закону

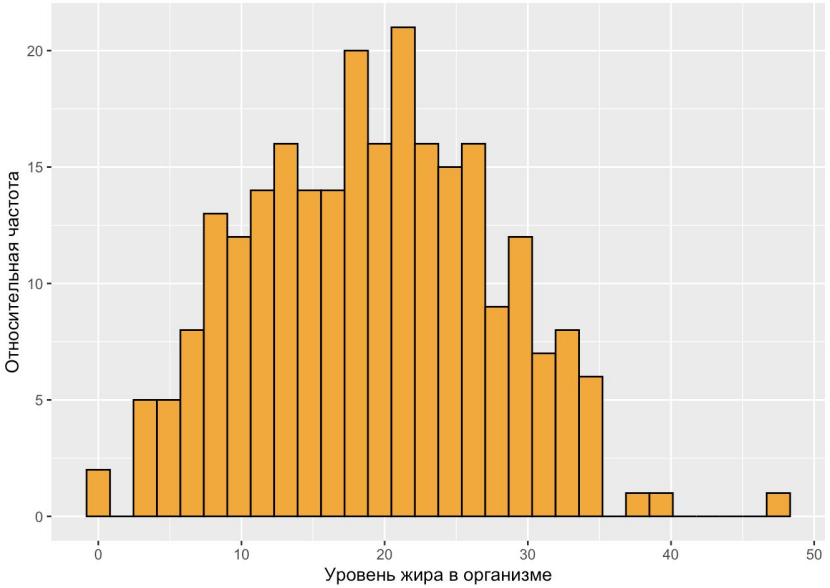


Рисунок 2. Гистограмма распределения значений переменной ‘BodyFat’ ДО удаления выброса.

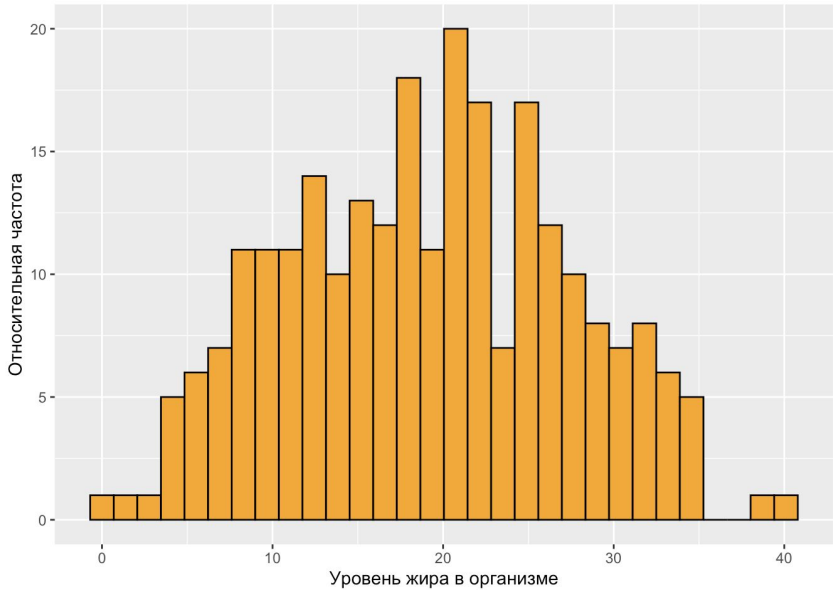


Рисунок 3. Гистограмма распределения значений переменной ‘BodyFat’ ПОСЛЕ удаления выброса

Коэффициент асимметрии = 0.145
Коэффициент эксцесса = -0.372

Таблица 3. Результаты теста на нормальность Шапиро-Уилка

Test statistic	P value
0.9917	0.1649

Корреляционный анализ

Различные типы зависимости переменных:

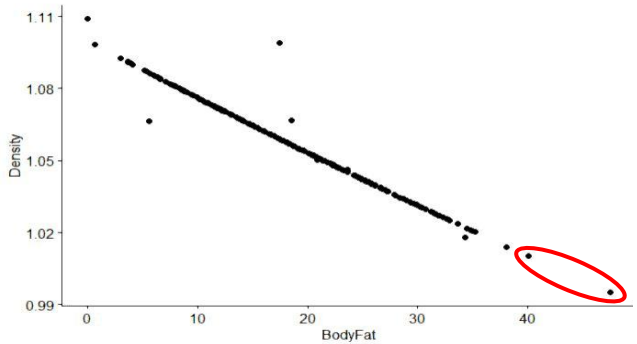


Рисунок 4. Диаграмма рассеивания для ‘BodyFat’ и ‘Density’.

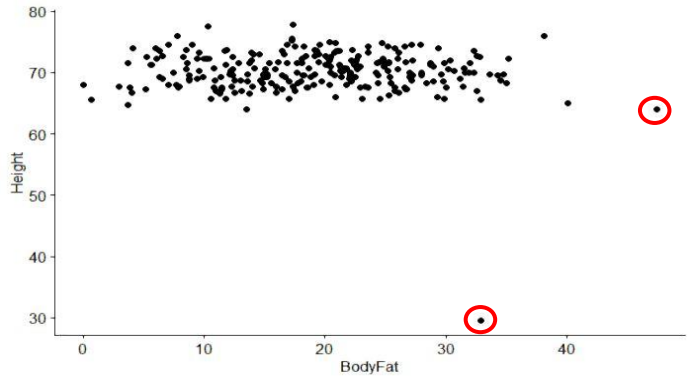


Рисунок 5. Диаграмма рассеивания для ‘BodyFat’ и ‘Height’.

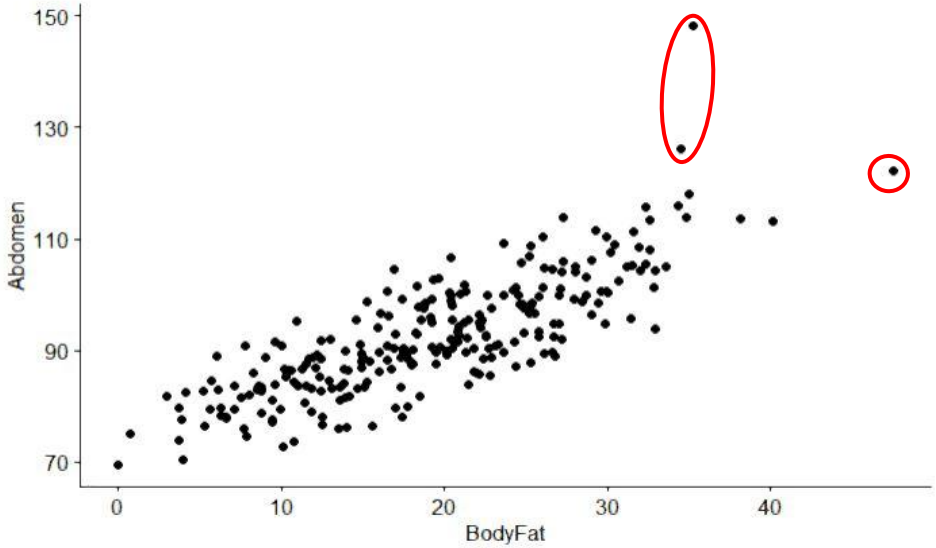


Рисунок 6. Диаграмма рассеивания для ‘BodyFat’ и ‘Abdomen’.

Значимость матрицы парных коэффициентов корреляции

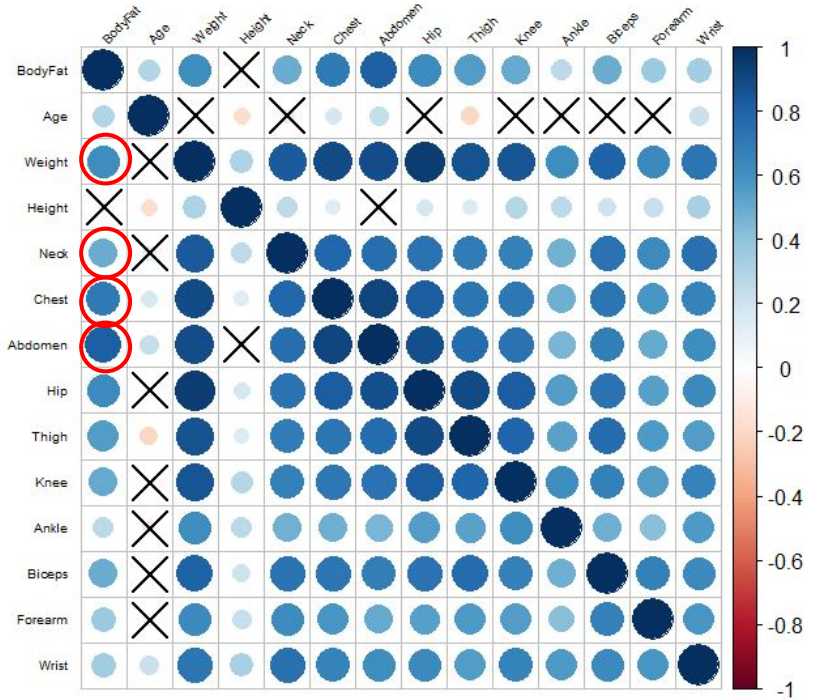


Рисунок 7. Корреляционная матрица (после удаления выбросов)

Некоторые коэффициенты корреляции изменились после удаления выбросов, следственно, **выбросы искажали** реальное значение парной корреляции между переменными. В целом, значения коэффициентов корреляции после удаления выбросов **изменились незначительно**.

С рассматриваемой переменной ‘BodyFat’ большинство признаков имеют **сильную прямую** взаимосвязь, признак Density же имеет **сильную обратную** (единственную обратную после удаления выбросов) взаимосвязь.

Матрица частных коэффициентов корреляции

	BodyFat	Age	Weight	Height	Neck
BodyFat	1.000000000	0.08463081	0.04087110	-0.018222726	-0.02663314
Age	0.084630812	1.00000000	-0.19447379	-0.125683376	0.10495551
Weight	0.040871101	-0.19447379	1.00000000	0.444742743	0.27233241
Height	-0.018222726	-0.12568338	0.44474274	1.000000000	-0.03112856
Neck	-0.026633141	0.10495551	0.27233241	-0.031128562	1.00000000
Chest	0.059138334	0.03276487	0.39722877	-0.213968186	0.02886588
Abdomen	0.037951787	0.29556499	0.23321620	-0.061834866	0.09762970
Hip	0.028656398	-0.07048095	0.51885438	-0.254728966	-0.18447252
Thigh	-0.025290681	-0.36989376	0.07117307	-0.196988552	0.11835960
Knee	-0.004207198	0.24861980	0.25363452	0.086331946	-0.14836610
Ankle	-0.084330733	-0.10349189	0.19384029	-0.059480699	-0.09463282
Biceps	-0.070118004	0.06763044	0.18826556	-0.046170579	0.09049183
Forearm	0.036922405	-0.17770765	0.02667392	-0.014009830	0.13101026
Wrist	0.002950963	0.34993119	0.11343952	0.088033788	0.25540480

Коэффициенты парной корреляции по модулю **значительно больше** частных коэффициентов, что говорит нам о том, что остальные переменные значительно **усиливают связь** между переменной ‘BodyFat’ и каждой ее парой.

Множественный коэффициент корреляции

```
Pearson's product-moment correlation

t = 105.68, df = 250,
p-value < 2.2e-16

95 percent confidence interval:
0.9859109 0.9914037

sample estimates:
      cor
  0.988993
```

Выводы:

Коэффициент значимый, существенный

‘BodyFat’ имеет очень тесную линейную корреляционную связь с другими переменными

Линейные регрессионные модели

Корреляционная матрица

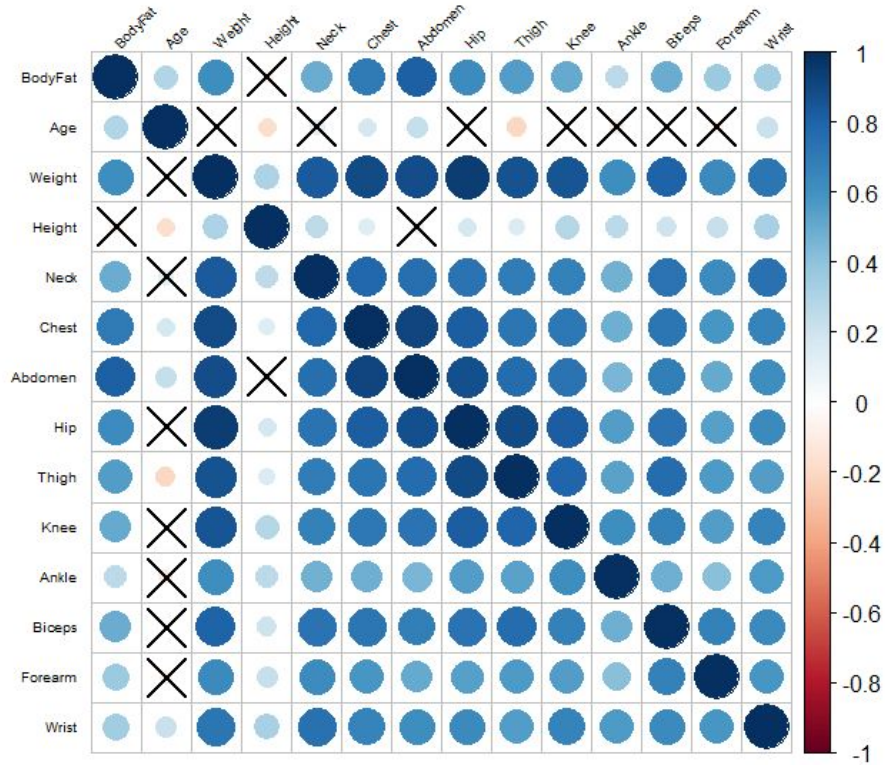


Рисунок 8. Корреляционная матрица.

Построение линейных регрессионных моделей

Первая линейная модель со всеми признаками:

$$y_{BodyFat} = -18.96 + 0.06637 \cdot x_{Age} - 0.08646 \cdot x_{Weight} - 0.06156 \cdot x_{Height} - 0.4815 \cdot x_{Neck} - 0.02335 \cdot x_{Chest} + 0.9349 \cdot x_{Abdomen} - 0.2273 \cdot x_{Hip} + 0.2618 \cdot x_{Thigh} +$$

$$+ 0.08228 \cdot x_{Knee} + 0.1542 \cdot x_{Ankle} + 0.1778 \cdot x_{Biceps} + 0.4532 \cdot x_{Forearm} - 1.607 \cdot x_{Wrist} + \varepsilon_i$$

Вторая линейная модель с признаками, отобранными по методу включения:

$$y_{BodyFat} = -21.54 + 0.07198 \cdot x_{Age} - 0.08279 \cdot x_{Weight} - 0.489 \cdot x_{Neck} + 0.9204 \cdot x_{Abdomen} - 0.2162 \cdot x_{Hip} + 0.3359 \cdot x_{Thigh} + 0.5175 \cdot x_{Forearm} - 1.511 \cdot x_{Wrist} + \varepsilon_i$$

Третья линейная модель со значимыми признаками:

$$y_{BodyFat} = -8.846 + 0.09553 \cdot x_{Age} - 0.6005 \cdot x_{Neck} + 0.7675 \cdot x_{Abdomen} + 0.5181 \cdot x_{Forearm} - 2.161 \cdot x_{Wrist} + \varepsilon_i$$

Сравнение линейных моделей

Таблица 4. Проверка допущений линейной регрессии.

Регрессионная модель	Линейная зависимость	Нормальность остатков	Независимость остатков	Гомоскедастичность остатков	Adjusted R^2
`lm1` (со всеми признаками)	✓	✓	✓	✓	72.6%
`lm2` (метод включения)	✓	✓	✓	✓	72.9%
`lm3` (со значимыми признаками)	✓	✓	✓	✓	71.6%

Таблица 5. Информационные критерии Акаике и Шварца (Баесовский инф. критерий).

Модель	Значение AIC	Значение BIC
lm1	1459.159	1512.041
lm2	1451.448	1486.703
lm3	1460.410	1485.088

Вывод

Лучшая модель - модель с признаками, отобранными по методу включения::

$$y_{BodyFat} = -21.54 + 0.07198 \cdot x_{Age} - 0.08279 \cdot x_{Weight} - 0.489 \cdot x_{Neck} + 0.9204 \cdot x_{Abdomen} - 0.2162 \cdot x_{Hip} + 0.3359 \cdot x_{Thigh} + 0.5175 \cdot x_{Forearm} - 1.511 \cdot x_{Wrist} + \varepsilon_i$$

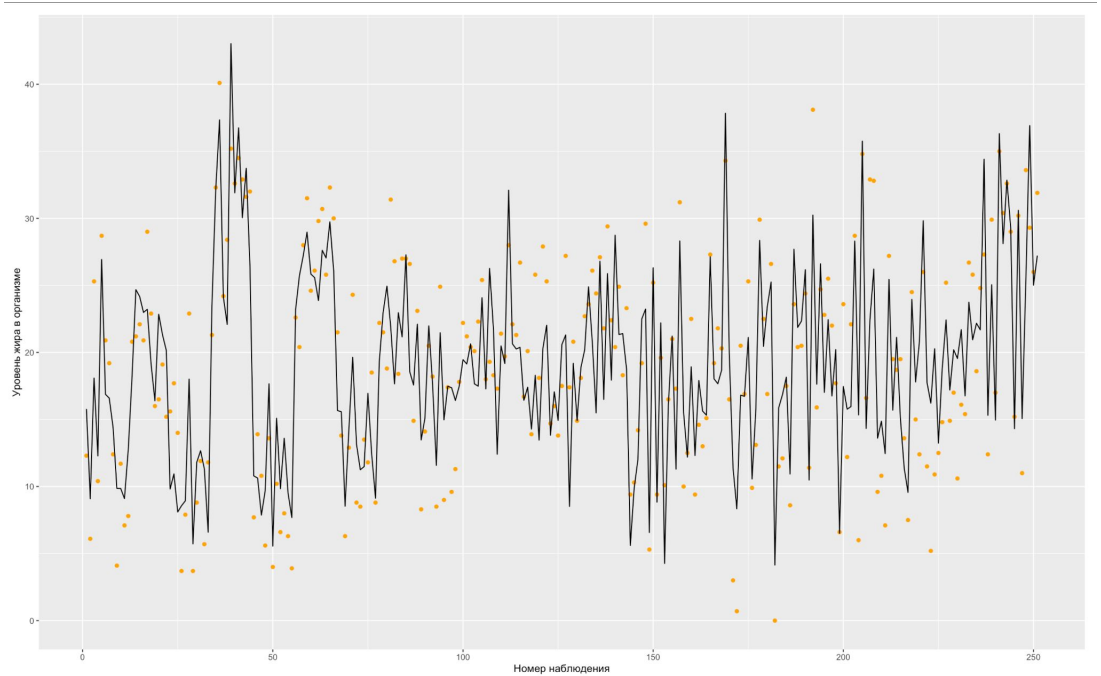


Рисунок 9. График наблюдаемых и модельных значений зависимой переменной

Регрессионный анализ. Нелинейная (степенная) регрессионная модель

Построение нелинейных регрессионных моделей

Экспоненциальная регрессионная модель:

$$y_{BodyFat} = e^{1.697} \cdot e^{0.006313 \cdot x_{Age} - 0.03676 \cdot x_{Neck} + 0.05187 \cdot x_{Abdomen} - 0.03366 \cdot x_{Hip} + 0.03748 \cdot x_{Thigh} - 0.07767 \cdot x_{Wrist} + \varepsilon_i}$$

Степенная регрессионная модель:

$$y_{BodyFat} = -6.068 \cdot x_{Age}^{0.137} \cdot x_{Weight}^{-0.4779} \cdot x_{Neck}^{-1.206} \cdot x_{Abdomen}^{4.787} \cdot x_{Hip}^{-2.116} \cdot x_{Thigh}^{1.307} \cdot x_{Forearm}^{0.696} \cdot x_{Wrist}^{-1.482} \cdot \varepsilon_i$$

Выбор лучшего нелинейного уравнения регрессии

Таблица 6. Таблица с характеристиками степенной модели с отобранными регрессорами.

Fitting linear model: BodyFat ~ Age + Weight + Neck + Abdomen + Hip + Thigh + Forearm + Wrist

Observations	Residual Std. Error	R^2	Adjusted R^2
244	0.2693	0.6663	0.655

Ни одна из построенных моделей не является адекватной, так как тесты на нормальность (тест Шапиро-Уилка и Жарка-Бера) распределения остатков отвергаются.

Выбор лучшего нелинейного уравнения регрессии

Таблица 7. Информационные критерии Акаике и Шварца (Баесовский инф. критерий).

Модель	Значение AIC	Значение BIC
nlm1	70.11073	122.56825
nlm2	68.17532	89.15833
nlm3	62.99759	97.96927

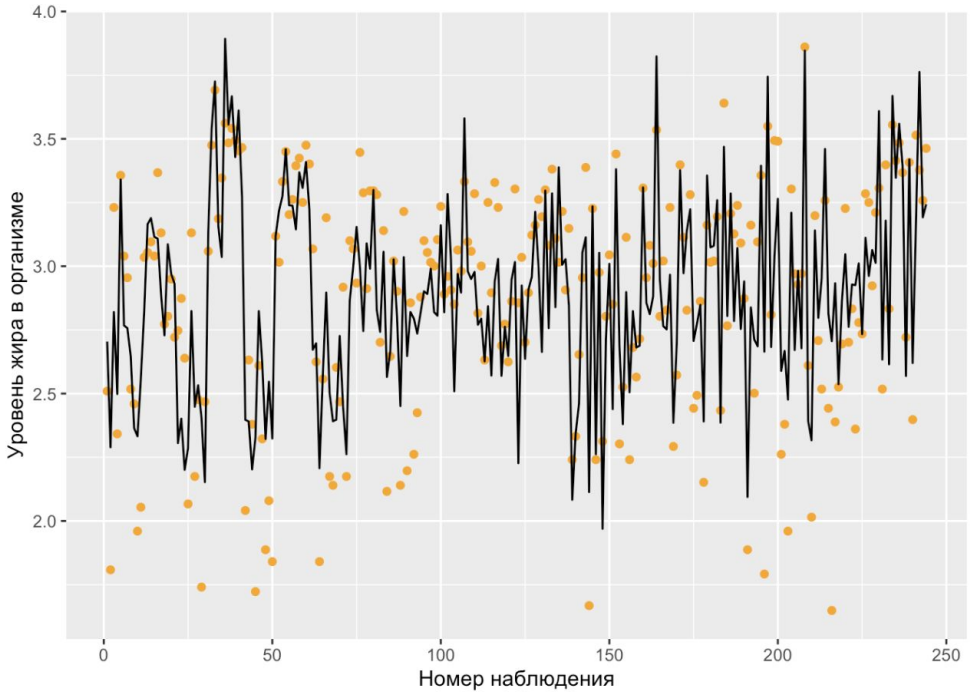


Рисунок 10. График наблюдаемых и модельных значений зависимой переменной

Интерпретация всех коэффициентов и характеристик

$$y_{BodyFat} = -6.068 \cdot x_{Age}^{0.137} \cdot x_{Weight}^{-0.4779} \cdot x_{Neck}^{-1.206} \cdot x_{Abdomen}^{4.787} \cdot x_{Hip}^{-2.116} \cdot x_{Thigh}^{1.307} \cdot x_{Forearm}^{0.696} \cdot x_{Wrist}^{-1.482} \cdot \varepsilon_i$$

```
## [1] "Эластичность Age: 0.179 %"
```

```
## [1] "Эластичность Weight: -0.858 %"
```

```
## [1] "Эластичность Neck: -1.52 %"
```

```
## [1] "Эластичность Abdomen: 7.511 %"
```

```
## [1] "Эластичность Hip: -3.377 %"
```

```
## [1] "Эластичность Thigh: 1.85 %"
```

```
## [1] "Эластичность Forearm: 0.809 %"
```

```
## [1] "Эластичность Wrist: -1.491 %"
```

Регрессионный анализ. Итог

Таблица 8. Проверка общих допущений у линейной и нелинейной регрессии.

Регрессионная модель	Нормальность остатков	Независимость остатков	Гомоскедастичность остатков
`lm3_1` (нелинейная)	✗	✓	✗
`lm2` (линейная)	✓	✓	✓

Победила линейная регрессия:

$$y_{BodyFat} = -21.54 + 0.07198 \cdot x_{Age} - 0.08279 \cdot x_{Weight} - 0.489 \cdot x_{Neck} + 0.9204 \cdot x_{Abdomen} - 0.2162 \cdot x_{Hip} + 0.3359 \cdot x_{Thigh} + 0.5175 \cdot x_{Forearm} - 1.511 \cdot x_{Wrist} + \varepsilon_i$$

Компьютерная работа №2

Выделение главных компонент (гк)

Предварительный анализ

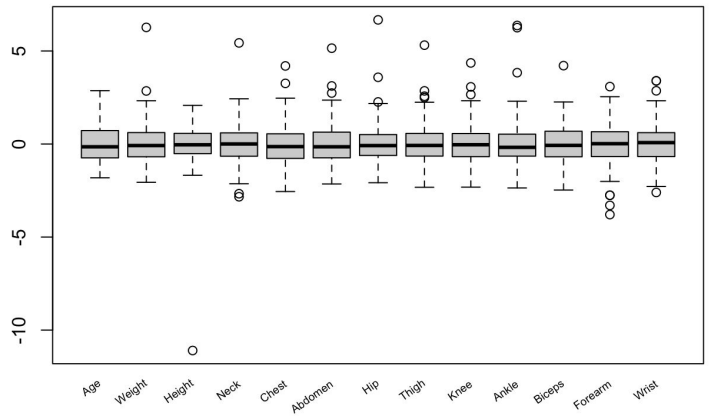


Рисунок 11. Ящичковые диаграммы отнормированных данных.

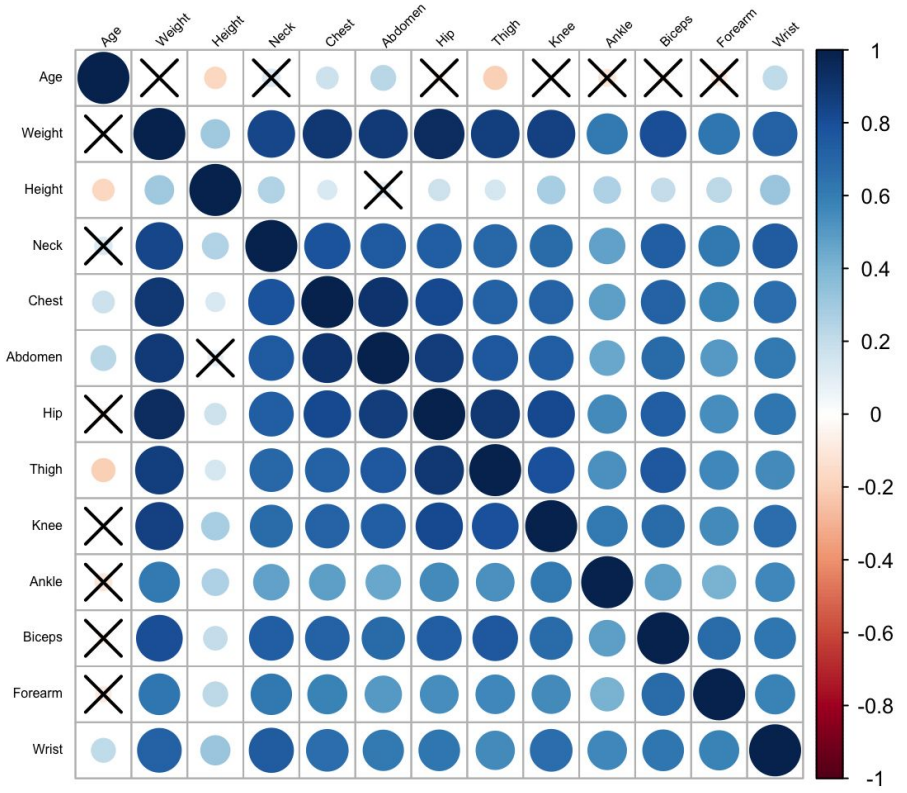


Рисунок 12. Корреляционная матрица.

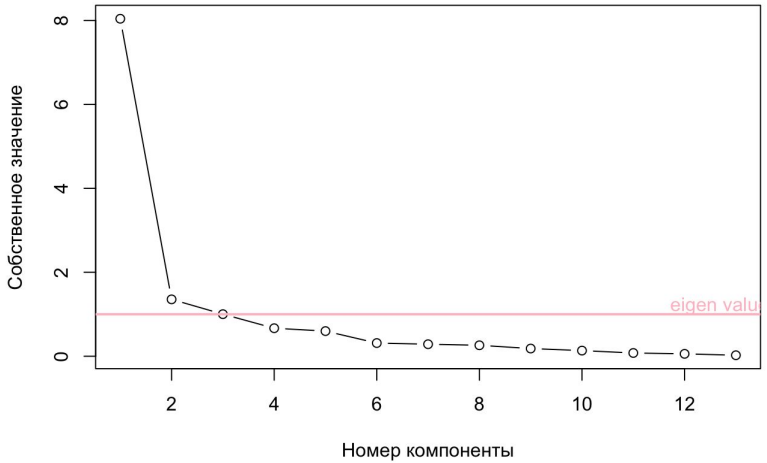


Рисунок 13. Выбор числа гк с помощью критерия Кайзера

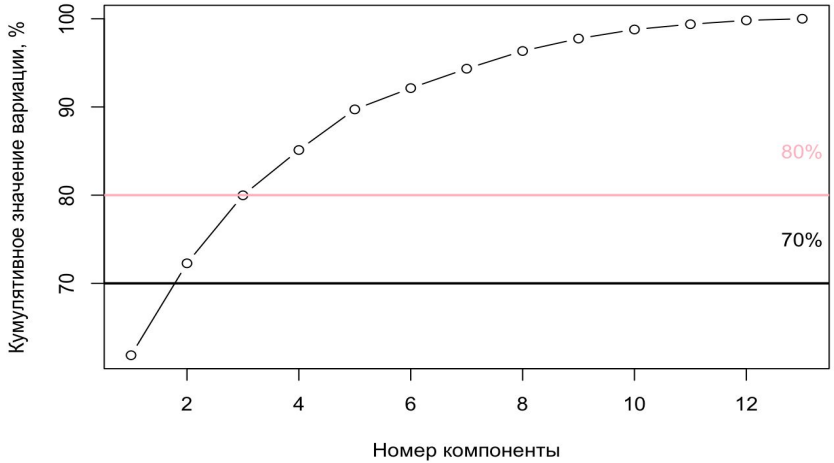


Рисунок 14. Выбор числа гк с помощью доли суммарной вариации

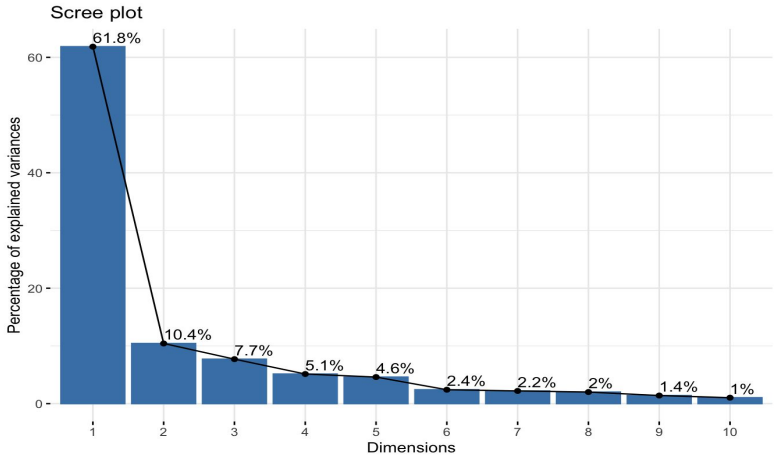


Рисунок 15. Выбор числа гк с помощью критерия каменистой осыпи

Результаты

Согласно проведенным тестам, оптимальное количество компонент - **3.**

Таблица 9. Матрица факторных нагрузок.

	PC1	PC2	PC3
Age	0.028	0.874	0.420
Weight	0.977	-0.021	-0.039
Height	0.287	-0.545	0.678
Neck	0.867	0.104	0.121
Chest	0.896	0.243	-0.061
Abdomen	0.884	0.308	-0.122
Hip	0.924	0.004	-0.221
Thigh	0.879	-0.143	-0.322
Knee	0.874	-0.058	0.001
Ankle	0.654	-0.261	0.128
Biceps	0.849	-0.057	-0.076
Forearm	0.708	-0.156	0.071
Wrist	0.791	0.094	0.388

Таблица 10. Матрица факторных нагрузок с варимакс-вращением.

	PC1	PC2	PC3
Age	0.003	-0.100	0.965
Weight	0.951	0.227	0.012
Height	0.063	0.909	-0.096
Neck	0.817	0.265	0.197
Chest	0.903	0.052	0.222
Abdomen	0.911	-0.035	0.245
Hip	0.946	0.052	-0.064
Thigh	0.915	0.034	-0.244
Knee	0.839	0.252	-0.003
Ankle	0.577	0.405	-0.122
Biceps	0.833	0.183	-0.044
Forearm	0.653	0.318	-0.059
Wrist	0.679	0.470	0.324

1) Измерения тела (Weight, Neck, Chest, Abdomen, Hip, Thigh, Knee, Biceps, Wrist, Ankle, Forearm, Wrist)
2) Рост (Height)
3) Возраст (Age)

Построение уравнения регрессии с использованием выделенных ГК

Построение уравнения регрессии с использованием выделенных ГК

Уравнение регрессии на ГК:

$$y_{BodyFat} = 19.15 + 1.829 \cdot Component_1 + 2.831 \cdot Component_2 - 1.718 \cdot Component_3 + \varepsilon_i$$

Таблица 11. Таблица с характеристиками уравнения регрессии с использованием выделенных ГК.

Observations	Residual Std. Error	R^2	Adjusted R^2
252	5.433	0.5836	0.5786

Данной моделью объясняется приблизительно **58% вариации зависимой переменной** исходя из значения Adjusted R^2. Также подмечаем, что **p-value всех компонент меньше 5%** уровня значимости, что еще раз подтверждает, что они являются значимыми для модели.

Построение уравнения регрессии с использованием выделенных ГК

Подчиняются ли остатки моделей нормальному распределению

Тест Харке-Бера

Test statistic	df	P value
1.186	2	0.5527

✔️

Проверка остатков на независимость

Теста Дарбина-Ватсона

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.1126364         1.7735   0.078
##  Alternative hypothesis: rho != 0
```

✔️

Проверка на гомоскедастичность

Тест Бреуша-Пагана

Test statistic	df	P value
48.54	3	1.638e-10 * * *

✔️

Построение уравнения регрессии с использованием выделенных ГК

Если сравнивать модели lm_1 и lm_GK, можно сделать предварительный вывод, что **первая модель лучше**, так как ее Adjusted R^2 выше.

Таблица 12. Сравнение регрессионных моделей.

Регрессионная модель	Нормальность остатков	Независимость остатков	Гомоскедастичность остатков	Adjusted R^2
`lm1` (линейная)	✓	✓	✓	72.9%
`lm1_1` (нелинейная)	×	✓	×	65.5%
`lm_GK` (на ГК)	✓	✓	✓	57.9%

Построение уравнения регрессии с использованием выделенных ГК

Для модели линейной регрессии lm1 информационные критерии Акаике и Шварца немного меньше, чем для модели lm_GK. Значит , **лучше использовать модель линейной регрессии из компьютерной работы 1.**

Таблица 13. Информационные критерии Акаике и Шварца (Баесовский инф. критерий).

Модель	Значение AIC	Значение BIC
lm1	1451.448	1486.703
lm_GK	1574.114	1591.761

Кластерный анализ

Определение оптимального числа кластеров

Таблица 14. Оптимальное число кластеров, полученное разными методами.

Метод локтя (elbow method)	2
Метод силуэтов (silhouettes)	2
Статистикой разрыва (Gap-статистика)	0

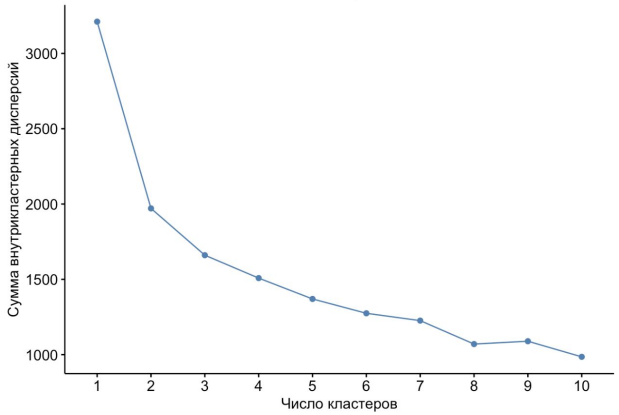


Рисунок 16. Зависимость WSS от числа кластеров.

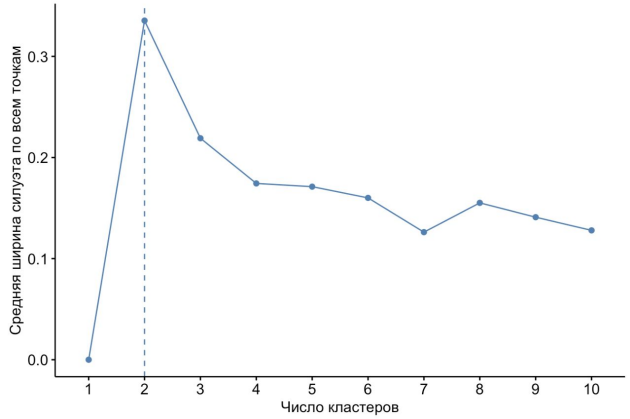


Рисунок 17. Зависимость средней ширины силуэта от числа кластеров.

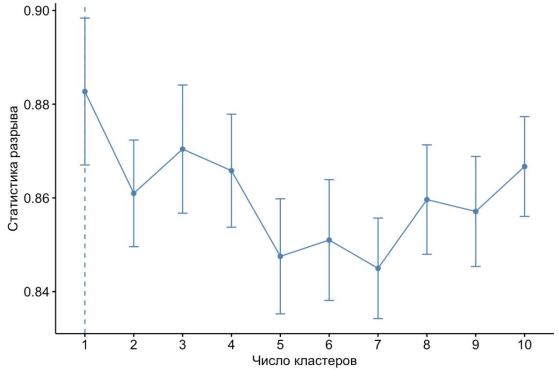
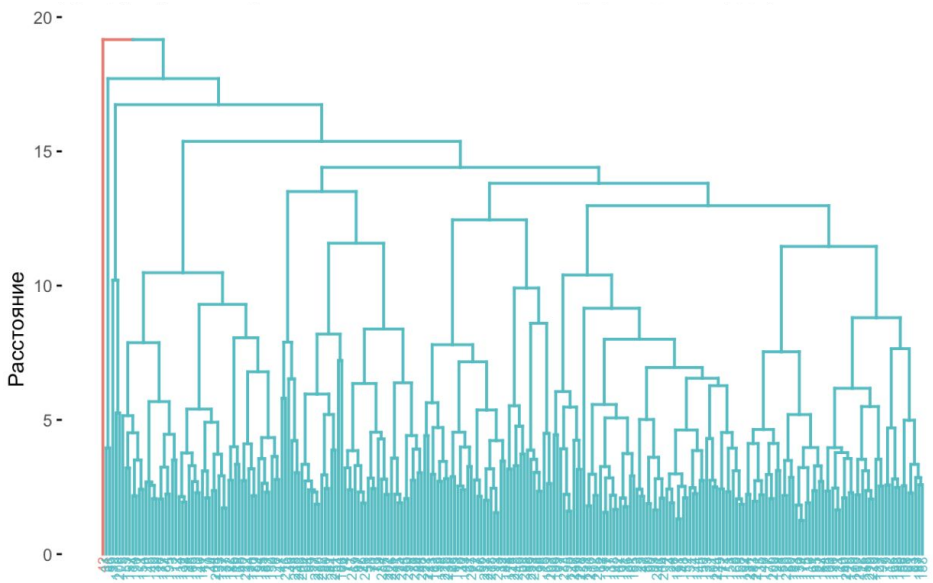


Рисунок 18. Зависимость статистики разрыва от числа кластеров.

Дендрограмма, расстояние Махаланобиса (принцип Варда)



Данные не подходят для проведения иерархической кластеризации.

Рисунок 19. Дендрограмма, расстояние Махаланобиса (принцип Варда).

Использование метода k-средних для классификации объектов

Метод k-средних

Размеры итоговых кластеров: 104, 144

Особенности кластеров:

- Признаки возраста и роста очень слабо влияют на всю классификацию, что подтверждается и корреляционной матрицей.
- Значения средних в кластерах по подавляющему большинству признаков находятся на примерно равном расстоянии, что говорит нам о схожих значениях различий по этим признакам в данных кластерах

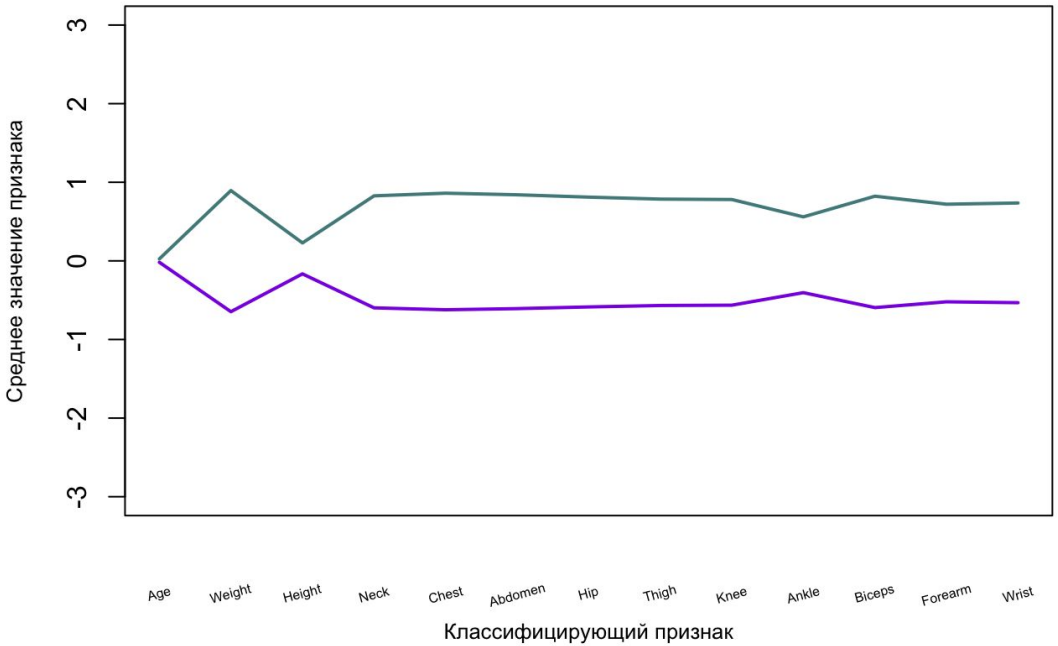


Рисунок 20. График средних (признаки стандартизированы).

Тест ANOVA

Практически во всех тестах p-value < 0.05, следовательно, с вероятностью ошибки 5% нулевая гипотеза о равенстве средних всех уровней фактора отвергается в пользу альтернативной гипотезы о том, что **между некоторыми уровнями есть различие в средних значениях**.

Поскольку у переменной “Age” p-value больше уровня значимости в 5%, то можем сказать, что между всеми уровнями фактора возраста **отсутствуют различия в средних значениях**.

Таблица 15. Результаты теста ANOVA между переменными ‘Age’ и ‘BodyFat’.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
as.factor(kmeans2$cluster)	1	0.1103	0.1103	0.1099	**0.7405**
Residuals	246	246.9	1.004	NA	NA

Деление на кластеры

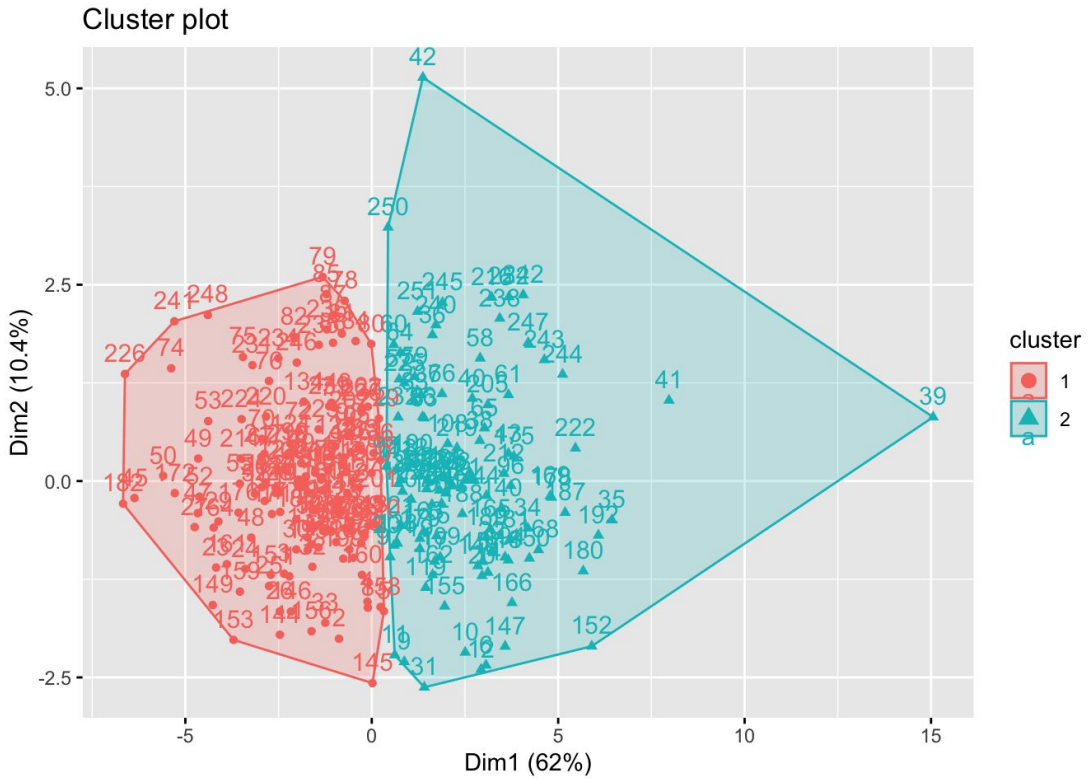


Рисунок 21. График деления наблюдений на кластеры.

Корреляция между кластерами и “BodyFat”:

[1] 0.5007077

“Процент жира” является ключевым обоснованием в делении на кластеры:

1 - “низкий процент жира" (144 набл.),
2- "высокий процент жира" (104 набл.)

Построение регрессионных моделей в кластерах (типологическая регрессия)

Построение уравнений регрессии в кластерах

Уравнение регрессии для 1 класса:

$$y_{BodyFat} = -26.9 + 0.07794 \cdot x_{Age} - 0.0315 \cdot x_{Weight} - 0.5515 \cdot x_{Neck} + 0.9105 \cdot x_{Abdomen} - 0.07732 \cdot x_{Hip} + 0.3111 \cdot x_{Thigh} + 0.507 \cdot x_{Forearm} - 2.159 \cdot x_{Wrist} + \varepsilon_i$$

Уравнение регрессии для 2 класса:

$$y_{BodyFat} = -24.63 + 0.06082 \cdot x_{Age} - 0.1154 \cdot x_{Weight} - 0.2704 \cdot x_{Neck} + 0.9564 \cdot x_{Abdomen} - 0.2222 \cdot x_{Hip} + 0.2803 \cdot x_{Thigh} + 0.4856 \cdot x_{Forearm} - 1.362 \cdot x_{Wrist} + \varepsilon_i$$

Таблица 16. Ранжирование признаков по убыванию силы воздействия на изменение результативного показателя в каждом классе.

Класс	1	2	3	4	5	6	7	8
1	Abdomen	Wrist	Neck	Thigh	Forearm	Hip	Weight	Age
2	Abdomen	Wrist	Weight	Hip	Thigh	Forearm	Neck	Age

Сопоставление качества построенных моделей

Таблица 17. Сравнение регрессионных моделей.

Регрессионная модель	Линейная зависимость	Нормальность остатков	Независимость остатков	Гомоскедастичность остатков	Adjusted R^2
`lm1` (для всей совокупности)	✓	✓	✓	✓	72.9%
`lm_class_1` (для 1-го класса)	✓	✓	×	✓	61.3%
`lm_class_2` (для 2-го класса)	✓	✓	✓	✓	70.2%

Итак, смотря на таблицу понимаем, что лучшая по качеству модель - линейная регрессионная модель для всей выборки.

$$y_{BodyFat} = -21.54 + 0.07198 \cdot x_{Age} - 0.08279 \cdot x_{Weight} - 0.489 \cdot x_{Neck} + 0.9204 \cdot x_{Abdomen} - 0.2162 \cdot x_{Hip} + 0.3359 \cdot x_{Thigh} + 0.5175 \cdot x_{Forearm} - 1.511 \cdot x_{Wrist} + \varepsilon_i$$

Дискриминантный анализ

Построение дискриминантных функций

Предварительно отбросив сильно (корр. >= 0.8) коррелированные признаки (Weight, Hip, Chest, Thigh) и используя результаты k-means, где было получено 2 кластера, мы приступили к построению дискриминантных функций

Априорные вероятности:

```
        1         2
0.5939394 0.4060606
```

Коэффициенты линейных дискриминантов:

```
                LD1
Age       0.05524701
Height    0.20712596
Neck      0.02775459
Abdomen   0.42173784
Knee      0.02919642
Ankle     0.27975834
Biceps    0.49498385
Forearm   0.32000801
Wrist     0.29189664
```

Классификационная матрица:

```
Classification table:
    obs
pred  1  2
   1 45  1
   2  2 35
Misclassification errors:
   1    2
4.26 2.78
[1] 3.52
```

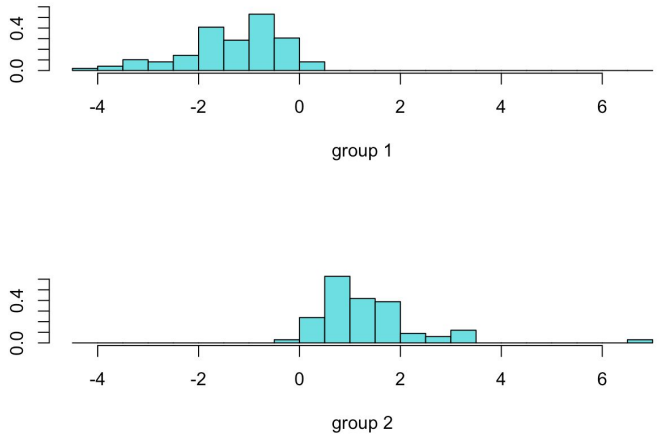


Рисунок 22. График значений дискриминантной функции на обучающей выборке.

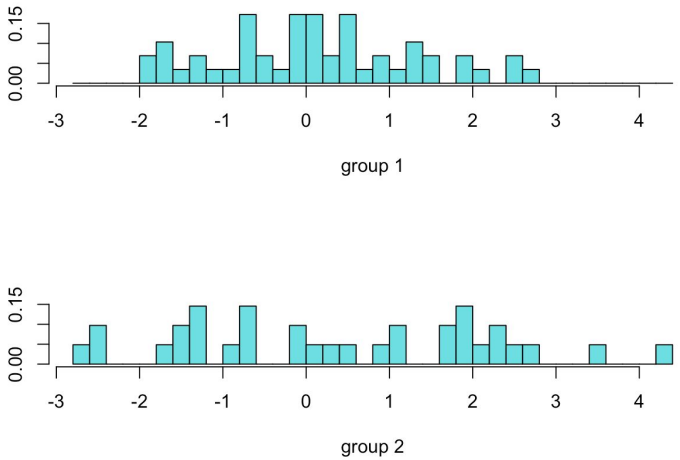


Рисунок 23. График предсказанных значений дискриминантной функции на тестовой выборке.

Таблица 18. Результаты проверки гипотезы о сходстве средних в исследуемых группах с помощью лямбды Уилкса.

	Df	Wilks	approx F	num Df	den Df	Pr(>F)
lda.pred$class	1	0.2888	19.97	9	73	2.021e-16
Residuals	81	NA	NA	NA	NA	NA

Отнесение новых объектов (3-4 наблюдения) к выделенным и описанным кластерам

Таблица 19. Характеристики отнесенных новых объектов к выделенным кластерам.

Номер наблюдения	Предсказанный класс	Апостериорная вероятность 1	Апостериорная вероятность 2	LD1	Реальное значение процента жира
68	1	0.96912433	0.030875667	-0.9023726	13.8
129	2	0.42908002	0.570919977	0.5002097	20.8
167	1	0.99081200	0.009188003	-1.3662137	21.8
249	2	0.01027686	0.989723137	2.1094548	33.6