МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

федеральное государственное бюджетное образовательное учреждение  
высшего образования

**«КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»**

**(ФГБОУ ВО «КубГУ»)**

**Факультет математики и компьютерных наук**

**Кафедра математических и компьютерных методов**

**КУРСОВАЯ РАБОТА**

**РЕГРЕССИОННЫЕ ЛИНЕЙНЫЕ МОДЕЛИ ДЛЯ ОПРЕДЕЛЕНИЯ ПОВТОРНОГО МАКСИМУМА ДЛЯ СИЛОВЫХ УПРАЖНЕНИЙ**

Работу выполнил \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Д. А. Пасько

(подпись, дата)

Курс 1

Направление подготовки 02.04.01 Математика и компьютерные науки

Направленность (профиль) Математическое и компьютерное   
 моделирование

Научный руководитель

докт. физ.-мат. наук,   
доцент \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ М. В. Голуб

(подпись, дата)

Нормоконтролер

ст. лаборант \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ И. С. Пшикова

(подпись, дата)

Краснодар

2020

**СОДЕРЖАНИЕ**

[ВВЕДЕНИЕ 3](#_Toc40367895)

[1 Зачем предсказывать повторные максимумы? 3](#_Toc40367896)

[2 О калькуляторах 4](#_Toc40367897)

[3 О данных и ресурсах 5](#_Toc40367898)

[4 Отправная точка 5](#_Toc40367899)

[1 Описание выборки и разведочный анализ данных 7](#_Toc40367900)

[1.1 О прошедших опрос 7](#_Toc40367901)

[1.2 Взаимодействия переменных 11](#_Toc40367902)

[1.3 Подведение итогов, отбор признаков 15](#_Toc40367903)

[2 Построение моделей 17](#_Toc40367904)

[2.1 Недостатки исходного решения 17](#_Toc40367905)

[2.2 Уточнение коэффициентов 20](#_Toc40367906)

[2.3 Уточнение коэффициентов: зависимость от диапазона 22](#_Toc40367907)

[2.4 Третье поколение моделей 25](#_Toc40367908)

[2.5 Сравнение линейных моделей и подведение итогов 30](#_Toc40367909)

[ЗАКЛЮЧЕНИЕ 35](#_Toc40367910)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 36](#_Toc40367911)

# 

# ВВЕДЕНИЕ

## 1 Зачем предсказывать повторные максимумы?

Очень часто при подготовке к соревнованиям по пауэрлифтингу или в процессе любительских тренировок полезно (а иногда и необходимо) оценить свои текущие возможности, не делая “проходку”, так как “проходка” оказывает сильное воздействие на нервную систему, её нежелательно делать часто и к ней нужно сначала подготавливаться, затем от неё отдыхать (чтобы суметь мобилизовать больше ресурсов и уменьшить риск травм), что занимает много времени и сил — и всё ради того, чтобы просто узнать, какой именно вес ты способен пожать, присесть или потянуть конкретно сейчас. Гораздо разумнее было бы оценить эти величины (*повторные максимумы*, ПМ), основываясь на результатах, которые ты показывал совсем недавно в процессе обычных тренировок или которые намного проще продемонстрировать (*многоповторные максимумы*, МПМ). Кроме того, если удастся построить такую модель, способную оценить силу человека, исходя из его последних достижений, можно будет сделать более явной связь между тренировками и реальными результатами, предотвращая перетренированность и временные потери на тренировки по программам, не дающим эффекта.

Итак. Зачем нужно предсказывать повторные максимумы? Регрессионные модели, которые требуется построить, помогут:

1) адекватно оценить свои возможности перед соревнованиями и заказать веса, очень близкие к реальному максимуму; кроме того, знать свой ПМ, не делая проходку, очень важно не только для тех, кто занимается пауэрлифтингом;

2) в период подготовки оценивать скорость своего прогресса (или вообще его наличие), предотвращая перетренированность; сравнивать свои результаты с результатами предыдущих циклов, даже если тогда использовалось другое число повторений;

3)более обоснованно планировать программу тренировок: пытаясь чередовать разные упражнения и разные диапазоны повторений, легко сделать программу, которая на самом деле не будет давать значимого эффекта или потребует неподъёмных усилий;

4) более точно замерять силовые качества групп мышц-антагонистов (очень редко для этого вообще возможно измерять ПМ у обеих групп, как и МПМ при одном и том же числе повторений);

5) выявить математические закономерности в силовых показателях человека.

## 2 О калькуляторах

К сожалению, многие люди отнеслись к этому исследованию очень скептически, поскольку уверены, что здесь имеется слишком сложная зависимость, чтобы её можно было описывать; другие утверждали, что знание такой информации принесёт больше вреда, чем пользы, если спортсмен не имеет большого опыта или не тренируется с тренером. Во многом они правы, но на самом деле, как я считаю, такое мнение обусловлено достаточно большим количеством разных калькуляторов и формул в Интернете, которые дают разные результаты и не ясно на чём основываются (на каких формулах либо откуда полученных).

Цель же этой работы – предоставить формулу, полученную в согласовании с концепциями машинного обучения, которая будет хорошо работать на многих людях, а не на десяти. Здесь написано, как она получена и почему.

## 3 О данных и ресурсах

Для сбора наблюдений был создан [опрос в Google Forms](https://forms.gle/R4zZqQJ3ggNdEuQ67) на русском языке и [его англоязычный аналог](https://forms.gle/dbthrvE5Y95beqUL9).

Русскоязычный опрос распространялся через соц. сеть «ВКонтакте» преимущественно в следующих группах: [Пауэрлифтинг | Тяжелая атлетика](https://vk.com/powerliftingnews), [Я ♥ ПАУЭРЛИФТИНГ](https://vk.com/ilovepowerlifting), [ПАУЭРЛИФТИНГ И ЖИМ ЛЁЖА 18+](https://vk.com/powerliftingworld), [Твой Тренер](https://vk.com/tvoytrenercom); аналогично англоязычный опрос распространялся на Facebook преимущественно здесь: [Powerlifting Motivation](https://www.facebook.com/PowerliftingMotivation/), [International Powerlifting League (IPL)](https://www.facebook.com/groups/IPLPowerlifting/), [Powerlifting Motivation Chat](https://www.facebook.com/groups/powerliftingmotivation/), [Powerlifting Memes](https://www.facebook.com/groups/powerliftingmemes/). Кроме того, несколько наблюдений я собрал лично.

Спустя два месяца от создания опросов данные были выгружены и началась их обработка. Теперь эти данные и всё, связанное с ними (в том числе последняя версия этого документа), хранятся [в моём репозитории](https://github.com/PasaOpasen/Powerlifting-training-diary-and-articles/tree/master/Estimating%20RM).

Для обработки данных и создания отчёта использовалась среда [RStudio](https://rstudio.com) и язык R версии 3.6.3.

## 4 Отправная точка

В книге [“Система тренировок КУБ”](https://forum.steelfactor.ru/index.php?showtopic=44596) на странице 23 (34 в оригинальной версии) приводится формула для определения повторного максимума:

где берётся некоторый рабочий вес и максимально возможное число повторений с ним. В оригинальных обозначениях:

К примеру, если вы можете пожать (лёжа) 100 кг на 5 раз, то из этой формулы следует, что вы сможете пожать 115 на раз и 90 на 9 раз (звучит правдоподобно).

Лично на мне эта формула хорошо работает и мне захотелось уточнить её для других людей разной комплекции и уровня подготовки. Также интересно то, что здесь имеется, по сути, очень простая линейная модель с двумя *предикторами*: рабочим весом и *взаимодействием* рабочего веса с числом повторений, и нет зависимости от самого упражнения, опыта человека и других характеристик, то есть в перспективе возможно с большой точностью описать наши возможности одной простой формулой, и это будет верно почти для всех людей, невзирая на все различия между ними.

# 

# 1 Описание выборки и разведочный анализ данных

## 1.1 О прошедших опрос

С помощью опроса было получено всего 174 достоверных наблюдения (исключая явные аномалии), принадлежащие предположительно 157 людям.

Все наблюдения содержат информацию о нескольких переменных. После некоторых преобразований над этими переменными получаем следующий набор признаков:

* *RM* – повторный максимум;
* *MRM* – многоповторный максимум;
* *Count* – количество повторений для многоповторного максимума;
* *Action* – движение для которого верны измерения (Жим, Тяга, Присед – самые базовые упражнения);
* *Sex* – пол испытуемого;
* *Experience* – группа опыта (До двух лет, 2-3 года, 4-5 лет, 6-10 лет, больше 10 лет);
* *Age* – возраст;
* *Weight* – собственный вес;
* *Height* – рост;
* *BodyType* – тип телосложения (Эктоморф, Мезоморф, Эндоморф). Типы телосложения в целом различаются скоростью обмена веществ и строением скелета, рычагами (хотя различия в строении скелета во многом являются следствием скорости обмена веществ);
* *CountGroup* – группа по диапазону повторений (2-3, 4-6, 7-10, 11-20, больше 20);
* *AgeGroup* – возрастная группа (меньше 20, 20-27, 28-35, больше 35);
* *Index* – [индекс массы тела](https://ru.wikipedia.org/wiki/Индекс_массы_тела), ИМТ;
* *IndexGroup* – группа по индексу массы тела (с градациями: выраженный дефицит, дефицит, норма, избыток, ожирение\_1, ожирение\_2, ожирение\_3 в соответствии с рекомендациями Всемирной Организации Здравоохранения).

Основные статистики по данным переменным представлены в таблице 1.

Таблица 1 – Основные статистики по опрошенным

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Переменная | mean | sd | median | mad | min | max | range |
| RM | 154.121 | 51.089 | 147.500 | 48.184 | 20.000 | 300.000 | 280.000 |
| MRM | 127.158 | 42.675 | 120.000 | 40.771 | 15.000 | 250.000 | 235.000 |
| Count | 8.144 | 7.071 | 6.000 | 4.448 | 2.000 | 35.000 | 33.000 |
| Age | 28.017 | 8.990 | 27.000 | 8.896 | 15.000 | 62.000 | 47.000 |
| Weight | 89.608 | 18.883 | 87.000 | 17.050 | 47.000 | 160.000 | 113.000 |
| Height | 176.931 | 8.252 | 177.000 | 8.896 | 155.000 | 198.000 | 43.000 |
| Index | 28.474 | 5.011 | 27.767 | 4.708 | 18.827 | 45.269 | 26.442 |

Сделаем некоторые выводы о выборке. Почти все испытуемые – мужчины (рисунок 1), поэтому *результаты, которые будут получены, не следует обобщать на женщин*. Возможно даже, что позднее придётся удалить принадлежащие женщинам наблюдения из выборки, если окажется, что те сильно выделяются.

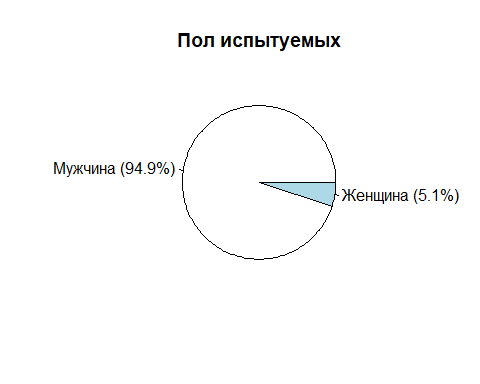


Рисунок 1 – пол испытуемых

Среди опрошенных почти половину составляли эндоморфы (рисунок 2).

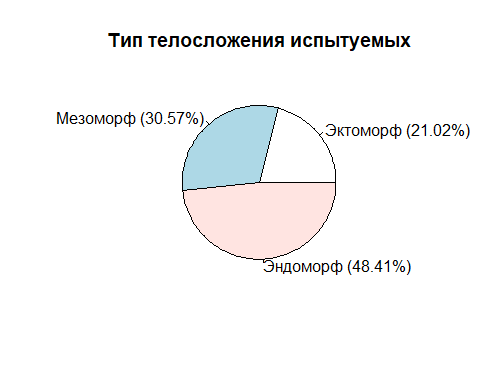


Рисунок 2 – тип телосложения испытуемых

Больше половины наблюдений относятся к жиму лёжа (рисунок 3). Я думаю, это связано с тем, что большинство спортсменов просто предпочитают это упражнение двум другим, вдобавок на жиме лёжа относительно проще замерять МПМ и, скорее всего, многие из ответивших специализировались именно на жиме лёжа.

Распределение наблюдений в группах по числу повторений меньше 11 – примерно равномерное (рисунок 4). Распределение одновременно по диапазону повторений и опыту тренировок представлено на рисунке 5.

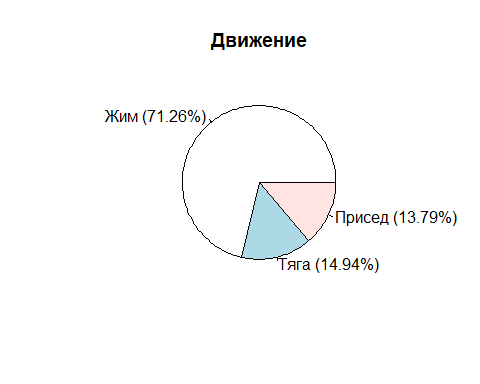


Рисунок 3 – тип упражнения

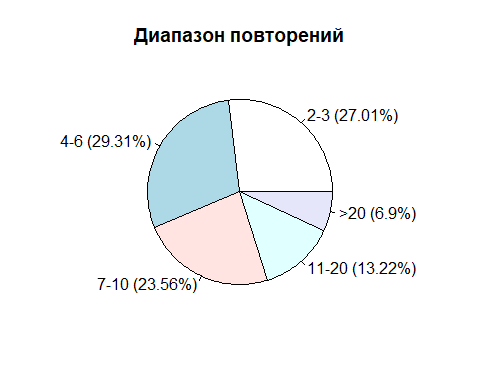


Рисунок 4 – диапазон повторений

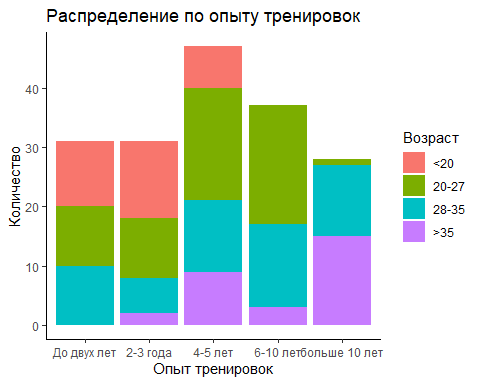


Рисунок 5 – распределение по опыту тренировок

## 1.2 Взаимодействия переменных

Иллюстрация некоторых парных взаимодействий между количественными признаками представлена на рисунке 7. Корреляции между ними – на рисунке 8. В таблице 2 представлены оценки коэффициента корреляции на каждое повторение от 2 до 10.

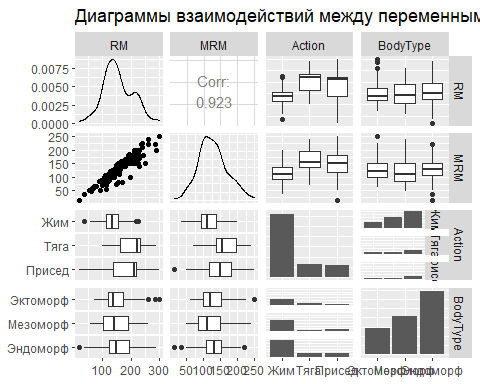


Рисунок 7 – диаграммы взаимодействий между переменными

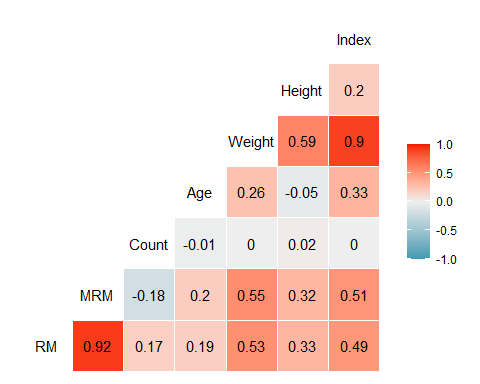


Рисунок 8 – корреляции между количественными переменными

Таблица 2 – Корреляция между MRM & RM для каждого повторения

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Число повторений | Корреляция MRM & RM | Нижняя граница | Верхняя граница | Всего наблюдений |
| 2 | 0.9980117 | 0.9949986 | 0.9992103 | 21 |
| 3 | 0.9919420 | 0.9818449 | 0.9964336 | 26 |
| 4 | 0.9947569 | 0.9838325 | 0.9983060 | 15 |
| 5 | 0.9891096 | 0.9734436 | 0.9955548 | 22 |
| 6 | 0.9942644 | 0.9814191 | 0.9982374 | 14 |
| 7 | 0.9936119 | 0.9687446 | 0.9987074 | 9 |
| 8 | 0.9964144 | 0.9876689 | 0.9989606 | 13 |
| 9 | 0.9581086 | 0.4902757 | 0.9973273 | 5 |
| 10 | 0.9890079 | 0.9645992 | 0.9966159 | 14 |

Есть ли разница в этих процентах для разных движений или телосложений? Дисперсионный анализ показывает, что нет, значимых различий не обнаружено (все p-значения больше 0.05) (таблица 3).

Таблица 3 – p-значения для разных факторных переменных на каждом числе повторений

|  |  |  |  |
| --- | --- | --- | --- |
| Число повторений | p-значения для телосложений | p-значения для типа движения | p-значения для групп по индексу массы |
| 2 | 0.9458770 | 0.7730952 | 0.8959581 |
| 3 | 0.8956176 | 0.1384093 | 0.2062476 |
| 4 | 0.9985364 | 0.279419 | 0.8345974 |
| 5 | 0.3427142 | 0.9919357 | 0.5228145 |
| 6 | 0.7269629 | 0.2637337 | 0.3154734 |
| 7 | 0.8204683 | мало данных | 0.9085377 |
| 8 | 0.6248755 | мало данных | 0.7177698 |
| 9 | 0.2564939 | мало данных | 0.4490970 |
| 10 | 0.0878330 | мало данных | 0.7676728 |

Применив тест Стьюдента для определения предсказываемых процентов и доверительных интервалов для них**,** получим результаты из таблицы 4.

Таблица 4 – Какой % составляет МПМ от ПМ

|  |  |  |  |
| --- | --- | --- | --- |
| Число повторений | Ожидаемый % | Нижняя граница | Верхняя граница |
| 2 | 93.54267 | 92.03110 | 95.05425 |
| 4 | 89.07451 | 86.50007 | 91.64895 |
| 6 | 85.58991 | 81.56746 | 89.61236 |
| 8 | 80.55848 | 78.43606 | 82.68090 |
| 10 | 77.84473 | 74.63339 | 81.05607 |
| 12 | 75.48574 | 68.66088 | 82.31060 |
| 15 | 74.47497 | 64.50324 | 84.44670 |
| 20 | 72.62619 | 54.19540 | 91.05698 |

В целом, эти данные согласуются с тем, что используются National Strength and Conditioning Association (NSCA) (<https://www.nsca.com>).

## 1.3 Подведение итогов, отбор признаков

Перечислим основные идеи об особенностях данных, сказанные ранее:

* Любые полученные результаты следует обобщать на женщин, так как в опросе они почти не участвовали. Возможно, ради точности вычислений даже придётся убрать женщин из выборки;
* Эндоморфы среди пауэрлифтеров встречаются чаще мезоморфов или эктоморфов. При этом у эндоморфов значимо чаще встречается “ожирение первой стадии”, чем у представителей других телосложений;
* Результаты в приседе и тяге растут с увеличением индекса массы тела лишь до некоторого порога. Увеличение индекса массы тела выше “ожирения первой стадии” не будет полезным;
* Сильная корреляция между индексом массы тела и повторным максимумом наблюдается у мезофорфов – в жиме, у эктоморфов – в тяге, у эндоморфов – в приседе;
* Не обнаружено значимой разницы значений для разных телосложений, движений, групп по индексу массы, опыта, возраста для любого конкретного числа повторений или диапазона повторений (но при этом обнаружилась разница между движениями для диапазонов повторений до 10); процентовки от National Strength and Conditioning Association (NSCA) с небольшими погрешностями верны для всех спортсменов и всех движений;
* Разбиение наблюдений на группы по диапазону повторений (который сделан согласовано с физиологическими соображениями) должно значительно повысить точность модели, причем для числа повторений до 10-20 эта модель крайне близка к линейной;
* Наблюдения свыше 20 повторений не будут учитываться, так как их мало, вдобавок этот диапазон слабо коррелирует с абсолютной силой;
* При более чем 10 повторений опытные спортсмены способны поднимать более близкие к максимальным веса, чем неопытные.

Соображения о предикторах для модели:

* (повторный максимум) обязательно зависит как от (многоповторный максимум) и (числа повторений), так и от – группы по числу повторений;
* может зависеть также от (типа телосложения) и (упражнения) либо от их бинарных модификаций (например, если людей разделить не на 3 телосложения, а на эндоморфов и нет); возможно, стоит также попробовать учесть индекс массы тела или собственно вес, имеющие корреляцию с не меньше 0.5.

# 2 Построение моделей

Вычислительные возможности позволяют создать и проверить огромное количество моделей. Цель исследователя – предложить несколько вариантов и выбрать среди них лучший. Кроме этого, нужно посмотреть, как работает модель на разных группах данных, чтобы обнаружить выбросы или обнаружить, что на такой-то группе модель вообще не может работать, из-за чего такую группу придётся исключить.

Итак, в модели представляют интерес следующие показатели:

* ошибки в разных группах данных;
* выбросы и влиятельные наблюдения;
* статистическая значимость модели;
* оценки кросс-валидации.

## 2.1 Недостатки исходного решения

Применив формулу

ко всем наблюдениям, получим оценки (Fact), которые будут отличаться от истинного (Target). Разница между этими величинами – остатки.

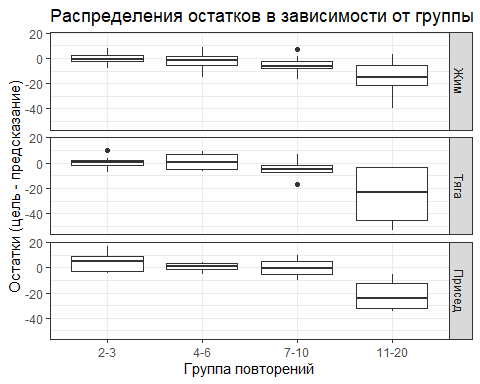


Рисунок 9 – распределение остатков в зависимости от группы для модели Вендлера

Здесь такая логика: чем лучше модель, тем “ящики” тоньше и тем их середины ближе к 0.

В данном случае видно, что модель Вендлера хорошо работает почти на всех данных из диапазона 2-3 и на немалой части данных из диапазона 4-6. Для большего числа повторений модель даёт завышенные оценки.

Если посмотреть на численные ошибки

## Наибольшая ошибка в большую сторону: 53.22   
## Наибольшая ошибка в меньшую сторону: 17.018   
## Модель ошиблась более чем на 5 % в 48 случаях из 162 ( 29.62963 %)  
## Модель ошиблась более чем на 5 кг в 69 случаях из 162 ( 42.59259 %)  
## ----Статистика по ошибкам в процентах:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.009091 1.466058 2.895000 4.466610 6.231771 24.950000   
## -------Среднеквадратичная ошибка: 10.79902

увидим, что модель заметно ошибается в 30% случаев и в среднем на почти 5%. Первые формулы [Мориса и Райдина](https://power-fitness.ru/metod-morisa-i-rajdina-ili-kak-uznat-svoj-maksimum-v-zhime-lezha.html) показывают худший результат:

## Наибольшая ошибка в большую сторону: 54.49607   
## Наибольшая ошибка в меньшую сторону: 40.32727   
## Модель ошиблась более чем на 5 % в 71 случаях из 162 ( 43.82716 %)  
## Модель ошиблась более чем на 5 кг в 92 случаях из 162 ( 56.79012 %)  
## ----Статистика по ошибкам в процентах:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.03358 1.89439 4.28278 5.81013 8.61114 34.06004   
## -------Среднеквадратичная ошибка: 13.73176

На самом деле это две формулы вида

где зависят от того, к какой части тела относится упражнение. Сама идея сделать логарифмическое преобразование используется нередко, но здесь это имеет негативный эффект: равен далеко не 0.99 (таблица 5), как сказано в первоисточнике.

Таблица 5 – Регрессионная модель по типу модели Мориса и Райдина

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr |
| **(Intercept)** | 0.06472 | 0.01162 | 5.57 | 1.071e-07 |
| **Action2Up** | -0.02185 | 0.01395 | -1.567 | 0.1191 |
| **Action2Down:Count** | 0.01701 | 0.001341 | 12.69 | 7.031e-26 |
| **Action2Up:Count** | 0.01823 | 0.001013 | 17.99 | 4.325e-40 |

Fitting linear model: log(RM/MRM) ~ Action2 + Count:Action2

|  |  |  |  |
| --- | --- | --- | --- |
| Observations | Residual Std. Error |  | Adjusted |
| 162 | 0.04448 | 0.7585 | 0.7539 |

## 2.2 Уточнение коэффициентов

Теперь возьмём модель вида и подберём коэффициенты лучшим образом. В итоге:

## Наибольшая ошибка в большую сторону: 19.33226   
## Наибольшая ошибка в меньшую сторону: 21.46997   
## Модель ошиблась более чем на 5 % в 39 случаях из 162 ( 24.07407 %)  
## Модель ошиблась более чем на 5 кг в 66 случаях из 162 ( 40.74074 %)  
## ----Статистика по ошибкам в процентах:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.001034 1.344302 3.072494 3.463881 4.816583 13.279030   
## -------Среднеквадратичная ошибка: 6.656834   
## Оценка кросс-валидации для всего набора данных 46.67453   
## Оценка кросс-валидации для не более чем 10 повторений 31.99091   
## Оценка кросс-валидации для не более чем 6 повторений 30.00755

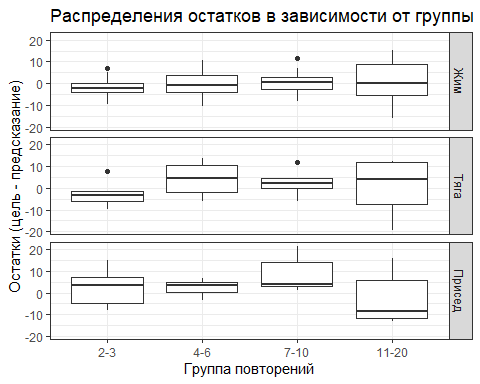


Рисунок 10 – распределение остатков в зависимости от группы для улучшенной модели

## ОБЩАЯ ИНФОРМАЦИЯ О МОДЕЛИ:  
##   
##   
## Call:  
## lm(formula = RM ~ MRM + MRM:Count - 1, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -19.3323 -4.3200 -0.4872 3.9362 21.4700   
##   
## Coefficients:  
## Estimate Std. Error t value Pr   
## MRM 1.0470734 0.0067757 154.53 меньше 2e-16 \*\*\*  
## MRM:Count 0.0209793 0.0008774 23.91 меньше 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.698 on 160 degrees of freedom  
## Multiple R-squared: 0.9983, Adjusted R-squared: 0.9983   
## F-statistic: 4.664e+04 on 2 and 160 DF, p-value: меньше 2.2e-16  
##   
##   
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS  
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:  
## Level of Significance = 0.05   
##   
## Call:  
## gvlma::gvlma(x = mdl)   
##   
## Value p-value Decision  
## Global Stat 15.1022 0.004494 Assumptions NOT satisfied!  
## Skewness 3.1467 0.076078 Assumptions acceptable.  
## Kurtosis 1.9628 0.161218 Assumptions acceptable.  
## Link Function 0.2725 0.601627 Assumptions acceptable.  
## Heteroscedasticity 9.7202 0.001823 Assumptions NOT satisfied!  
## Value p-value Decision  
## Global Stat 15.1022329 0.004493809 Assumptions NOT satisfied!  
## Skewness 3.1467476 0.076078463 Assumptions acceptable.  
## Kurtosis 1.9627685 0.161217544 Assumptions acceptable.  
## Link Function 0.2725495 0.601626657 Assumptions acceptable.  
## Heteroscedasticity 9.7201673 0.001822569 Assumptions NOT satisfied!

##   
## ТЕСТ НА НОРМАЛЬНОСТЬ РАСПРЕДЕЛЕНИЯ ОСТАТКОВ  
##   
##   
## Shapiro-Wilk normality test  
##   
## data: mdl$residuals  
## W = 0.98782, p-value = 0.1726

Выше приведена вся необходимая информация о качестве модели (назовём эту модель ). Из неё важнее всего следующее:

* На выборочных данных модель ошибается максимум на 21 кг и 13%;
* В среднем модель ошибается на 3%;
* В 75% случаев ошибка не превышала 5%;
* Наибольший разброс ошибок приходится на диапазон 11-20;
* Модель статистически значима и удовлетворяет всем нужным требованиям, кроме гетероскедастичности;
* Немало больших ошибок приходится на диапазон 11-20 повторений, содержащий меньше всего наблюдений.

## 2.3 Уточнение коэффициентов: зависимость от диапазона

Теперь сделаем поправку для коэффициентов в зависимости от факторных переменных.

Путём подбора удалось найти две схожие модели (назовём их и ), немного отличающиеся оценками кросс-валидации на разных диапазонах. Поскольку на диапазоне повторений 11-20 по-прежнему сохранялись сильные ошибки, было принято решение удалить этот диапазон (также это можно обосновать тем, что на этом диапазоне меньше данных и сложнее отлавливать выбросы, да и физиологически он не так тесно связан с силой, как другие, о чём уже было сказано).

Далее приведён анализ для модели , в которой и и имеют поправку на диапазон повторений.

## Наибольшая ошибка в большую сторону: 12.46227   
## Наибольшая ошибка в меньшую сторону: 18.24883   
## Модель ошиблась более чем на 5 % в 28 случаях из 139 ( 20.14388 %)  
## Модель ошиблась более чем на 5 кг в 44 случаях из 139 ( 31.65468 %)  
## ----Статистика по ошибкам в процентах:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00977 0.91712 2.53843 2.98739 4.29788 12.47021   
## -------Среднеквадратичная ошибка: 5.529521   
## Оценка кросс-валидации для всего набора данных 34.03878   
## Оценка кросс-валидации для не более чем 10 повторений 34.20106   
## Оценка кросс-валидации для не более чем 6 повторений 29.84518

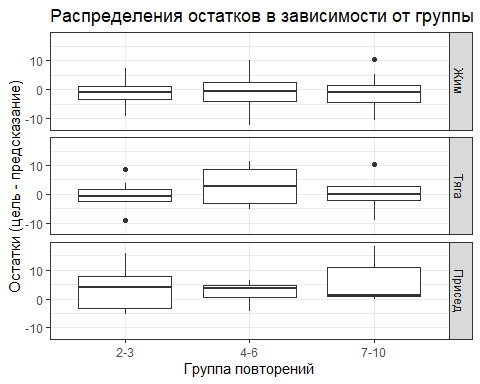


Рисунок 11 – распределение остатков в зависимости от группы для второй улучшенной модели

## ОБЩАЯ ИНФОРМАЦИЯ О МОДЕЛИ:  
##   
##   
## Call:  
## lm(formula = RM ~ MRM:Count:CountGroup + MRM:CountGroup - 1,   
## data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.4623 -3.6210 -0.1014 2.5825 18.2488   
##   
## Coefficients:  
## Estimate Std. Error t value Pr   
## MRM:CountGroup2-3 0.988699 0.029669 33.324 меньше 2e-16 \*\*\*  
## MRM:CountGroup4-6 0.997533 0.035651 27.981 меньше 2e-16 \*\*\*  
## MRM:CountGroup7-10 1.039799 0.055896 18.603 меньше 2e-16 \*\*\*  
## MRM:Count:CountGroup2-3 0.039572 0.011331 3.492 0.000650 \*\*\*  
## MRM:Count:CountGroup4-6 0.031609 0.007065 4.474 1.63e-05 \*\*\*  
## MRM:Count:CountGroup7-10 0.023854 0.006434 3.708 0.000306 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.653 on 133 degrees of freedom  
## Multiple R-squared: 0.9987, Adjusted R-squared: 0.9987   
## F-statistic: 1.768e+04 on 6 and 133 DF, p-value: меньше 2.2e-16  
##   
##   
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS  
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:  
## Level of Significance = 0.05   
##   
## Call:  
## gvlma::gvlma(x = mdl)   
##   
## Value p-value Decision  
## Global Stat 8.359 0.07929 Assumptions acceptable.  
## Skewness 2.795 0.09455 Assumptions acceptable.  
## Kurtosis 2.394 0.12180 Assumptions acceptable.  
## Link Function -2.689 1.00000 Assumptions acceptable.  
## Heteroscedasticity 5.858 0.01550 Assumptions NOT satisfied!  
## Value p-value Decision  
## Global Stat 8.358530 0.07929330 Assumptions acceptable.  
## Skewness 2.795168 0.09454886 Assumptions acceptable.  
## Kurtosis 2.394055 0.12179736 Assumptions acceptable.  
## Link Function -2.689003 1.00000000 Assumptions acceptable.  
## Heteroscedasticity 5.858310 0.01550367 Assumptions NOT satisfied!

##   
## ТЕСТ НА НОРМАЛЬНОСТЬ РАСПРЕДЕЛЕНИЯ ОСТАТКОВ  
##   
##   
## Shapiro-Wilk normality test  
##   
## data: mdl$residuals  
## W = 0.98086, p-value = 0.0484

## ФАКТОР ИНФЛЯЦИИ ДИСПЕРСИЙ:  
##   
## GVIF Df GVIF^(1/(2\*Df))  
## MRM:CountGroup 60763.96 3 6.270098  
## MRM:Count:CountGroup 60763.96 3 6.270098

Аналогичные выводы:

* На выборочных данных модель ошибается максимум на 18 кг и 13%;
* В среднем модель ошибается на 3%;
* В 80% случаев ошибка не превышала 5%;
* Наибольший разброс ошибок приходится на диапазон 4-6;
* Модель статистически значима и удовлетворяет всем нужным требованиям, кроме гетероскедастичности.

## 2.4 Третье поколение моделей

Путём пошагового отбора переменных были найдены ещё две равнозначные модели ( и ), которые, несмотря на б**о**льшую сложность в сравнении с предыдущими, имеют лучшие оценки при кросс-валидации.

Статистика для модели :

## Наибольшая ошибка в большую сторону: 12.09735   
## Наибольшая ошибка в меньшую сторону: 13.90043   
## Модель ошиблась более чем на 5 % в 21 случаях из 139 ( 15.10791 %)  
## Модель ошиблась более чем на 5 кг в 44 случаях из 139 ( 31.65468 %)  
## ----Статистика по ошибкам в процентах:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.01759 0.94961 2.08660 2.83833 3.94832 12.31595   
## -------Среднеквадратичная ошибка: 5.026615   
## Оценка кросс-валидации для всего набора данных 31.61373   
## Оценка кросс-валидации для не более чем 10 повторений 29.42087   
## Оценка кросс-валидации для не более чем 6 повторений 29.97883

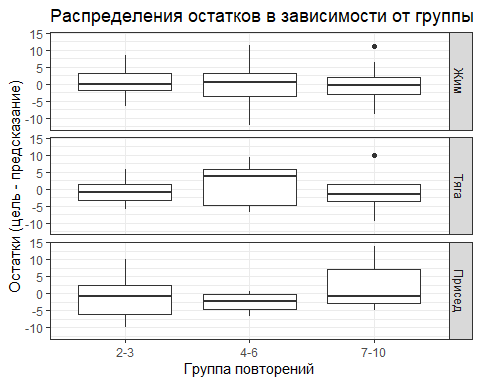


Рисунок 12 – распределение остатков в зависимости от группы для лучшей модели

## ОБЩАЯ ИНФОРМАЦИЯ О МОДЕЛИ:  
##   
##   
## Call:  
## lm(formula = RM ~ I((MRM/Index)^6) + MRM:CountGroup + MRM:Action +   
## MRM:CountGroup:Count - 1, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.0974 -3.2898 -0.1568 3.0781 13.9004   
##   
## Coefficients:  
## Estimate Std. Error t value Pr   
## I((MRM/Index)^6) -2.989e-05 1.125e-05 -2.657 0.00887 \*\*   
## MRM:CountGroup2-3 9.963e-01 2.848e-02 34.984 меньше 2e-16 \*\*\*  
## MRM:CountGroup4-6 9.711e-01 3.401e-02 28.558 меньше 2e-16 \*\*\*  
## MRM:CountGroup7-10 1.077e+00 5.195e-02 20.738 меньше 2e-16 \*\*\*  
## MRM:ActionТяга 2.893e-02 1.014e-02 2.854 0.00502 \*\*   
## MRM:ActionПрисед 4.984e-02 9.648e-03 5.166 8.78e-07 \*\*\*  
## MRM:CountGroup2-3:Count 3.325e-02 1.067e-02 3.116 0.00226 \*\*   
## MRM:CountGroup4-6:Count 3.578e-02 6.842e-03 5.229 6.62e-07 \*\*\*  
## MRM:CountGroup7-10:Count 1.833e-02 6.011e-03 3.050 0.00278 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.198 on 130 degrees of freedom  
## Multiple R-squared: 0.999, Adjusted R-squared: 0.9989   
## F-statistic: 1.395e+04 on 9 and 130 DF, p-value: меньше 2.2e-16  
##   
##   
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS  
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:  
## Level of Significance = 0.05   
##   
## Call:  
## gvlma::gvlma(x = mdl)   
##   
## Value p-value Decision  
## Global Stat 2.19263 0.7004 Assumptions acceptable.  
## Skewness 0.36972 0.5432 Assumptions acceptable.  
## Kurtosis 0.02639 0.8710 Assumptions acceptable.  
## Link Function 1.38357 0.2395 Assumptions acceptable.  
## Heteroscedasticity 0.41295 0.5205 Assumptions acceptable.  
## Value p-value Decision  
## Global Stat 2.19262765 0.7003792 Assumptions acceptable.  
## Skewness 0.36971780 0.5431581 Assumptions acceptable.  
## Kurtosis 0.02639011 0.8709512 Assumptions acceptable.  
## Link Function 1.38356953 0.2394941 Assumptions acceptable.  
## Heteroscedasticity 0.41295021 0.5204759 Assumptions acceptable.  
##

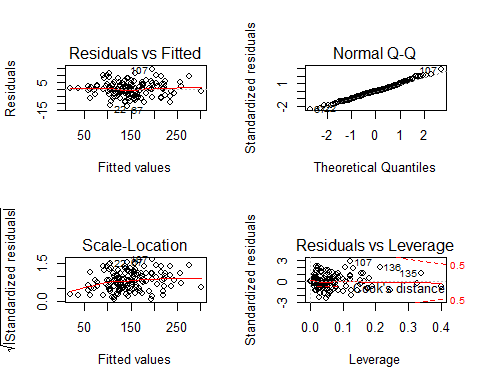


Рисунок 13 – статистика по лучшей модели

##   
## ТЕСТ НА НОРМАЛЬНОСТЬ РАСПРЕДЕЛЕНИЯ ОСТАТКОВ  
##   
##   
## Shapiro-Wilk normality test  
##   
## data: mdl$residuals  
## W = 0.99366, p-value = 0.7982

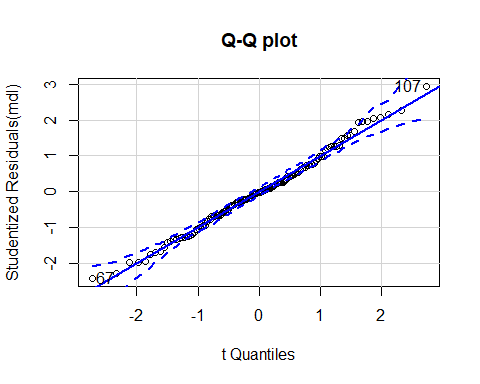


Рисунок 14 – qqplot для лучшей модели

## ФАКТОР ИНФЛЯЦИИ ДИСПЕРСИЙ:  
##   
## GVIF Df GVIF^(1/(2\*Df))  
## I((MRM/Index)^6) 2.561915 1 1.600598  
## MRM:CountGroup 72443.114733 3 6.456535  
## MRM:Action 2.898866 2 1.304839  
## MRM:CountGroup:Count 72727.913065 3 6.460759

Аналогичные выводы:

* На выборочных данных модель ошибается максимум на 13 кг и 12%;
* В среднем модель ошибается на менее 3%;
* В 85% случаев ошибка не превышала 5%;
* В целом, на каждом диапазоне имеется почти одинаковый разброс;
* Модель статистически значима и удовлетворяет всем нужным требованиям.

## 2.5 Сравнение линейных моделей и подведение итогов

В результате перекрёстной проверки для всех моделей при числе блоков 5, 6, 7, 8, 9, 10, 11, 12 было обнаружено, что модель значительно превосходит остальные модели, причём это верно, если делать проверку как на всём диапазоне повторений (2-10), так и на более близком к силовому диапазоне 2-7:

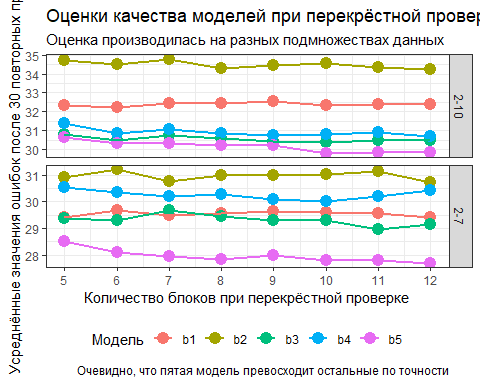


Рисунок 14 – ошибки при перекрестной проверке для лучших представленных моделей

На втором месте после модели идёт более простая модель .

Что это за модели? Посмотрим на модель :

##   
## Call:  
## lm(formula = RM ~ MRM:Count:CountGroup + MRM:Action - 1, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.1658 -3.3898 0.0727 3.0208 13.5361   
##   
## Coefficients:  
## Estimate Std. Error t value Pr   
## MRM:ActionЖим 0.998895 0.019934 50.111 меньше 2e-16 \*\*\*  
## MRM:ActionТяга 1.012918 0.020963 48.320 меньше 2e-16 \*\*\*  
## MRM:ActionПрисед 1.036228 0.021859 47.406 меньше 2e-16 \*\*\*  
## MRM:Count:CountGroup2-3 0.031222 0.007776 4.015 9.86e-05 \*\*\*  
## MRM:Count:CountGroup4-6 0.029581 0.004048 7.308 2.26e-11 \*\*\*  
## MRM:Count:CountGroup7-10 0.027353 0.002420 11.301 меньше 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.31 on 133 degrees of freedom  
## Multiple R-squared: 0.9989, Adjusted R-squared: 0.9988   
## F-statistic: 2.004e+04 on 6 and 133 DF, p-value: меньше 2.2e-16

Она имеет вид

где равен 0.9988949 для жима, 1.0129181 для тяги и 1.0362281 для приседа, а есть поправка на диапазон повторений, равная 0.0312223 для диапазона 2-3, 0.0295811 – для 4-6 и 0.0273531 – для 7-10. Как видно, первые три коэффициента близки к единице, однако различия между ними имеют значение; заметим также, что поправка на диапазон повторений 2-3 довольно близка к коэффициенту Вендлера (0.0333).

Кстати, по точности модель превосходит вторую модель Мориса (ту, что выражена таблицами для каждого движения):

## Наибольшая ошибка в большую сторону: 25.1   
## Наибольшая ошибка в меньшую сторону: 20.6   
## Модель ошиблась более чем на 5 % в 60 случаях из 139 ( 43.16547 %)  
## Модель ошиблась более чем на 5 кг в 80 случаях из 139 ( 57.55396 %)  
## ----Статистика по ошибкам в процентах:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 2.006 4.325 5.282 7.386 19.308   
## -------Среднеквадратичная ошибка: 9.482539

## Наибольшая ошибка в большую сторону: 11.16581   
## Наибольшая ошибка в меньшую сторону: 13.53614   
## Модель ошиблась более чем на 5 % в 21 случаях из 139 ( 15.10791 %)  
## Модель ошиблась более чем на 5 кг в 44 случаях из 139 ( 31.65468 %)  
## ----Статистика по ошибкам в процентах:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.02251 1.03594 2.41179 2.89557 4.13407 11.44670   
## -------Среднеквадратичная ошибка: 5.194366

Тогда аналогичная таблица, на сколько нужно умножить свой вес для разного числа повторений, выглядит так:

Таблица 6 – На сколько нужно умножить свой вес для разного числа повторений, чтобы получить 1ПМ

|  |  |  |  |
| --- | --- | --- | --- |
| Повторы | Присед | Жим | Тяга |
| 1 | 1.000000 | 1.000000 | 1.000000 |
| 2 | 1.098673 | 1.061340 | 1.075363 |
| 3 | 1.129895 | 1.092562 | 1.106585 |
| 4 | 1.154553 | 1.117220 | 1.131243 |
| 5 | 1.184134 | 1.146801 | 1.160824 |
| 6 | 1.213715 | 1.176382 | 1.190405 |
| 7 | 1.227700 | 1.190367 | 1.204390 |
| 8 | 1.255053 | 1.217720 | 1.231743 |
| 9 | 1.282406 | 1.245073 | 1.259096 |
| 10 | 1.309759 | 1.272426 | 1.286449 |

Модель имеет два дополнительных слагаемых:

##   
## Call:  
## lm(formula = RM ~ I((MRM/Index)^6) + MRM:CountGroup + MRM:Action +   
## MRM:CountGroup:Count - 1, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.0974 -3.2898 -0.1568 3.0781 13.9004   
##   
## Coefficients:  
## Estimate Std. Error t value Pr   
## I((MRM/Index)^6) -2.989e-05 1.125e-05 -2.657 0.00887 \*\*   
## MRM:CountGroup2-3 9.963e-01 2.848e-02 34.984 меньше 2e-16 \*\*\*  
## MRM:CountGroup4-6 9.711e-01 3.401e-02 28.558 меньше 2e-16 \*\*\*  
## MRM:CountGroup7-10 1.077e+00 5.195e-02 20.738 меньше 2e-16 \*\*\*  
## MRM:ActionТяга 2.893e-02 1.014e-02 2.854 0.00502 \*\*   
## MRM:ActionПрисед 4.984e-02 9.648e-03 5.166 8.78e-07 \*\*\*  
## MRM:CountGroup2-3:Count 3.325e-02 1.067e-02 3.116 0.00226 \*\*   
## MRM:CountGroup4-6:Count 3.578e-02 6.842e-03 5.229 6.62e-07 \*\*\*  
## MRM:CountGroup7-10:Count 1.833e-02 6.011e-03 3.050 0.00278 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.198 on 130 degrees of freedom  
## Multiple R-squared: 0.999, Adjusted R-squared: 0.9989   
## F-statistic: 1.395e+04 on 9 and 130 DF, p-value: меньше 2.2e-16

Её вид, аналогично:

Здесь при последнем слагаемом равен -0.0000299, – индекс массы тела.

Очень интересно, что в этих двух моделях нет зависимости от возраста, типа телосложения, опыта и пола. Конечно, нельзя рассмотреть всевозможные вариации таких моделей, но среди сотни рассмотренных (в том числе с помощью пакета caret) не было обнаружено доказательств значимости включения этих факторов или их взаимодействий в модель. Может быть, для обнаружения подобной зависимости здесь недостаточно данных.

Как итог, для определения ПМ нужны лишь следующие данные:

* МПМ;
* Число повторений;
* Движение;
* Рост и вес спортсмена.

# ЗАКЛЮЧЕНИЕ

В результате проведённого исследования была найдена модель, не зависящая от типа телосложения, опыта тренировок и многих других характеристик, но при этом дающая хорошие результаты на 85% данных.

Я уверен, что большие погрешности в единичных случаях – это та часть закономерности, зависящая от человека и многих неучтённых характеристик. Тем не менее, влияние этих неучтённых факторов далеко не так велико, как предполагали многие: намного больше повторный максимум зависит от обычной математики.

Было продемонстрировано, что для числа повторений выше 10 нельзя построить достаточно точную модель (по крайней мере, при текущем наборе данных), но для диапазона 2-10 – ещё как можно.

Диапазон выше 10 повторений был отсеян по следующим причинам:

1) физиологически менее сильная связь между ПМ и большим числом повторений;

2) меньшая детерминированность между ПМ и большим числом повторений, обнаруженная при разведочном анализе данных, отчего трудно выявить достоверность данных и построить достаточно точную модель;

3) маленький объём данных, разбросанных очень неравномерно на большом диапазоне (от 11 до 35).

Посчитать собственные результаты при помощи найденной модели можно по ссылке <https://dmitrypasko.shinyapps.io/RMbyMRMestimating/>.

# 

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Хэтфилд Ф. К., Всестороннее руководство по развитию силы // Красноярск: Союзспорт, 1992. - 284 с.

2 Уикем Х., Гроулмунд Г. Язык R в задачах науки о данных: импорт, обработка, визуализация и моделирование данных. : Пер. с англ. — СПб. : ООО “Диалектика”, 2018. — 592 с.

3 Мастицкий С. Э., Шитиков В. К. Статистический анализ и визуализация данных с помощью R. – М.: ДМК Пресс, 2015. – 496 с.

4 Роберт И. Кабаков. R в действии. Анализ и визуализация данных в программе R. / пер. с англ. Полины А. Волковой. – М: ДМК Пресс, 2016. – 588 с.

5 Джеймс Г., Уиттон Дм Хасти Т., Тибширани Р. Введение в статистическое обучение с примерами на языке R. Пер. с англ. С. Э. Мастицкого - М.: ДМК Пресс, 2016. - 450 с:

6 Крупкина, Т. В. Математическая статистика [Электронный ресурс] : курс лекций / Т. В. Крупкина, А. К. Гречкосеев. – Электрон. дан. (3 Мб). – Красноярск : ИПК СФУ, 2009.