Отчёт к лабораторной работе

по дисциплине  
«Интеллектуальный анализ данных»

выполнил   
студент гр. ИС/б-18-1-з Демиденко А. А.  
зачётная книжка № 481483  
принял Шумейко И. П.

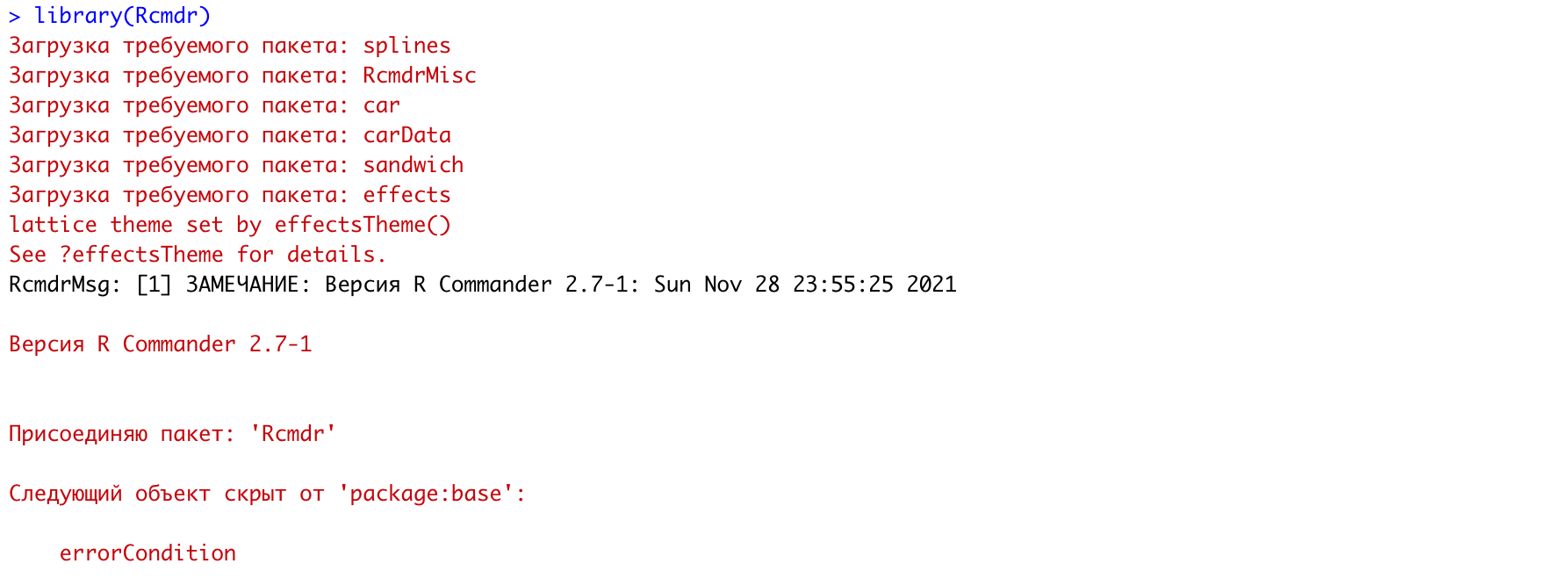
Лабораторной работа № 2.4  
«Корреляционный и регрессионный анализ данных. Множественная линейная регрессия»

## Цель работы

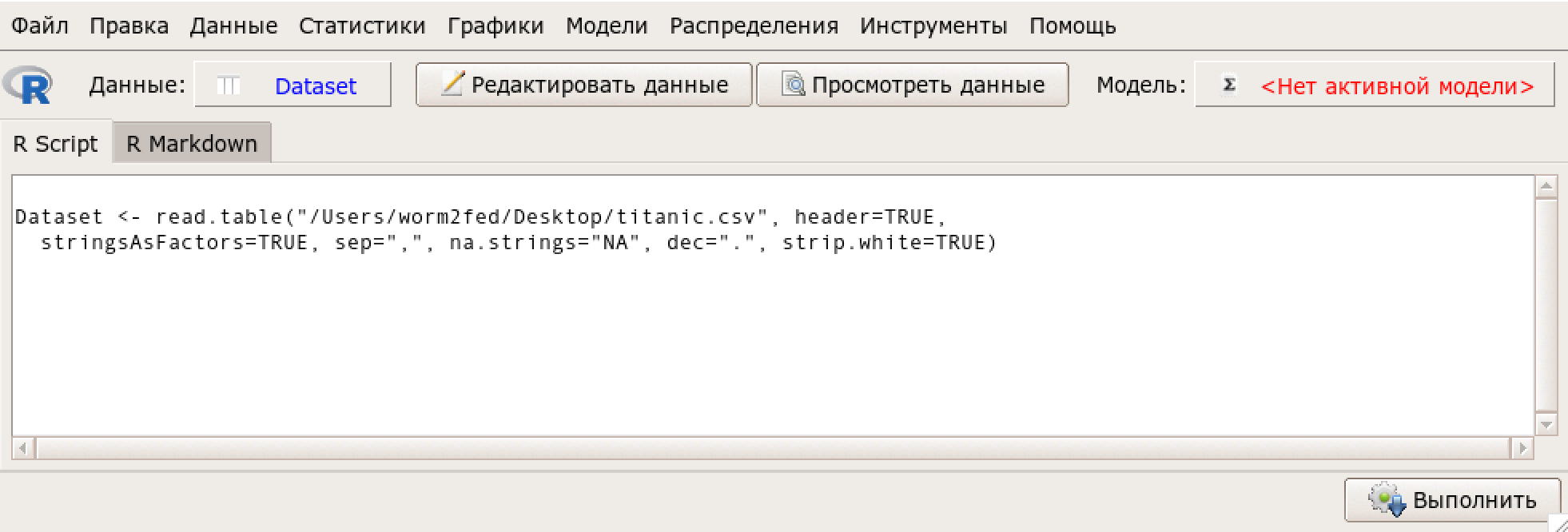
* исследовать возможности языка R для построения множественной линейной регрессии

## Ход работы

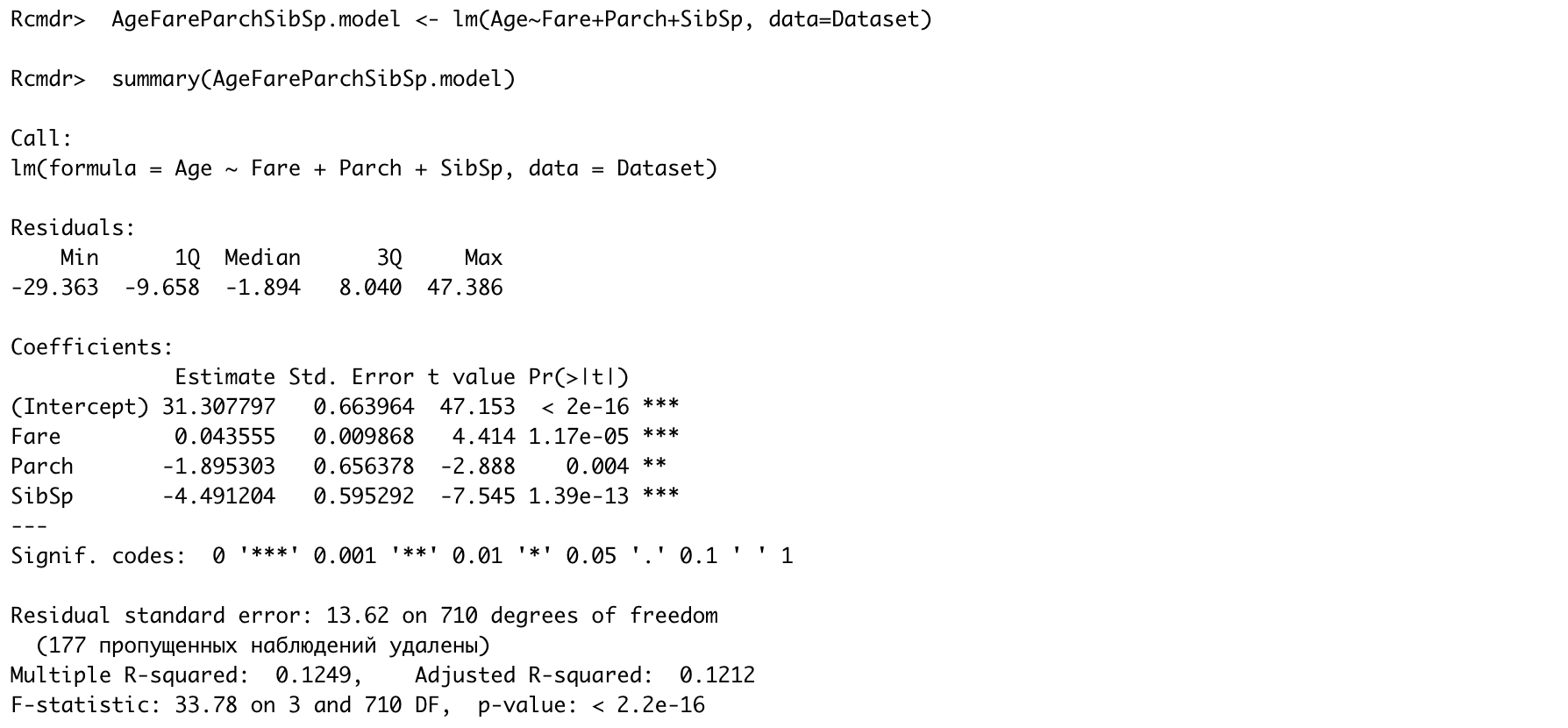
1. Запустим пакет R commander (Rcmdr):



Загрузим экспериментальные данные для анализа из файла (см. рисунок 1):

  
Рисунок 1 – Импорт данных для анализа

Построим множественную линейную регрессию. За x (исходные данные) возьмём цену билета, количество родственников первого и второго порядка, а за y (предсказываемая величина) – возраст.

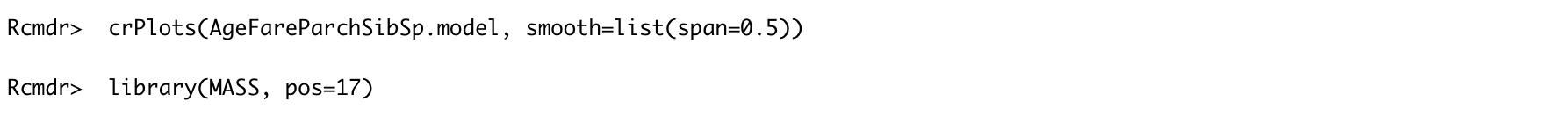


По таблице коэффициентов запишем полученное уравнение регрессии:

По коэффициенту R-squared определим, что модель описывает данные на 12%.

Далее смотрим на значение p-value. В нашем случае оно равно < 2.2e-16, то есть предсказываемая величина значительно зависит от исходных данных.

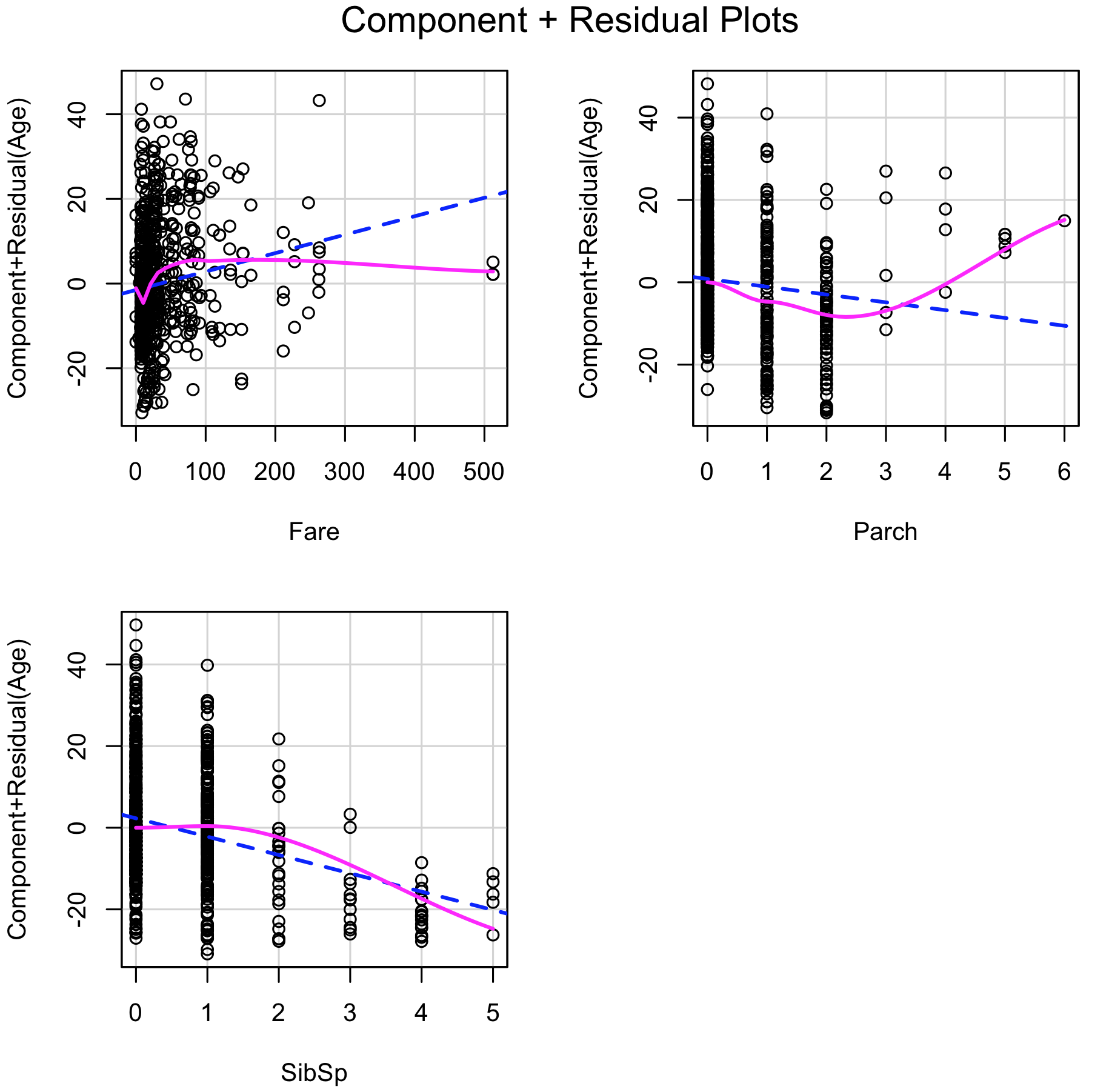
Построим графики остатков (см. рисунок 2).



На графиках видно, что наиболее точно модель предсказывает отношение возраста и количество родственников второго порядка (третий график).

Отношение возраста и количества родственников первого порядка (второй график) неплохо иллюстрирует взаимосвязь до 3 родственников включительно, а после – ведёт себя непредсказуемо.

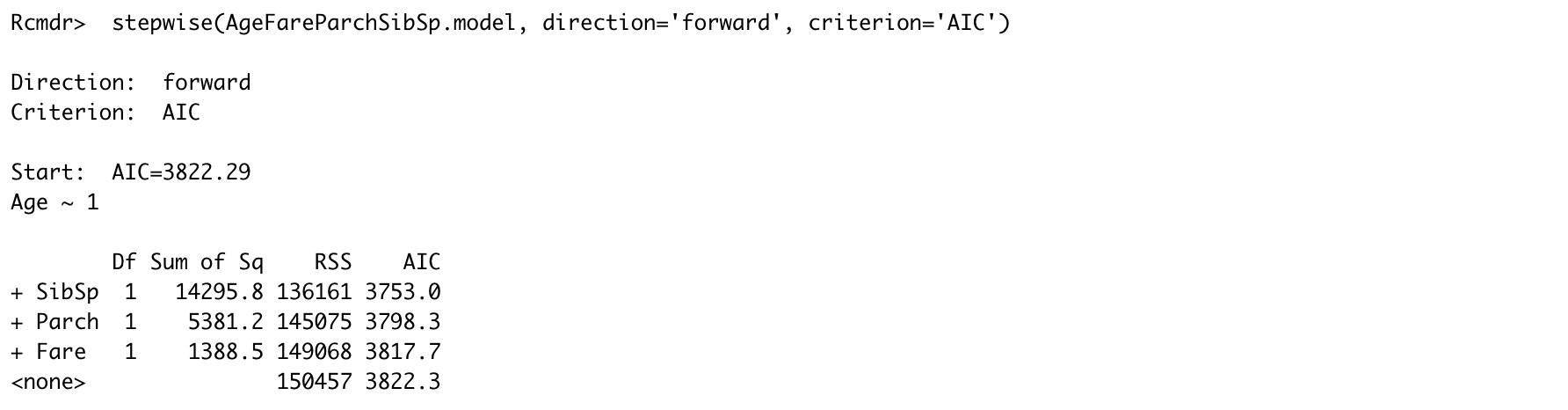
Ну а отношение возраста и цены билета вообще соотносятся очень плохо, как будто бы они мало связаны.

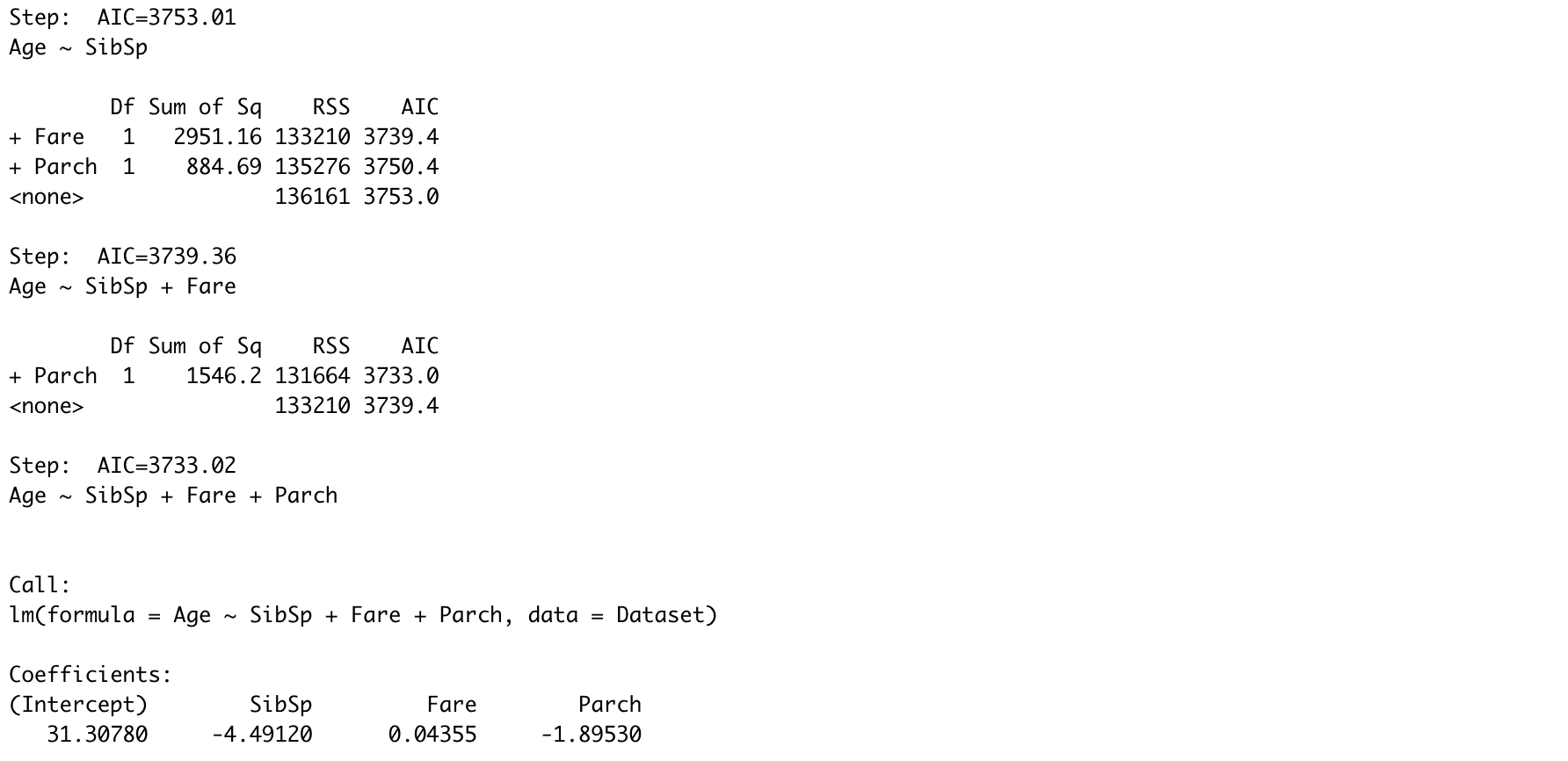
  
Рисунок 2 – Множественная линейная регрессия

Произведём пошаговую линейную регрессию с нашими экспериментальными данными.

Для начала выберем прямую пошаговую регрессию (метод прямого включения), заключающейся в том, что первоначально строится модель с одной экзогенной переменной, затем добавляется следующая и строится новая модель.

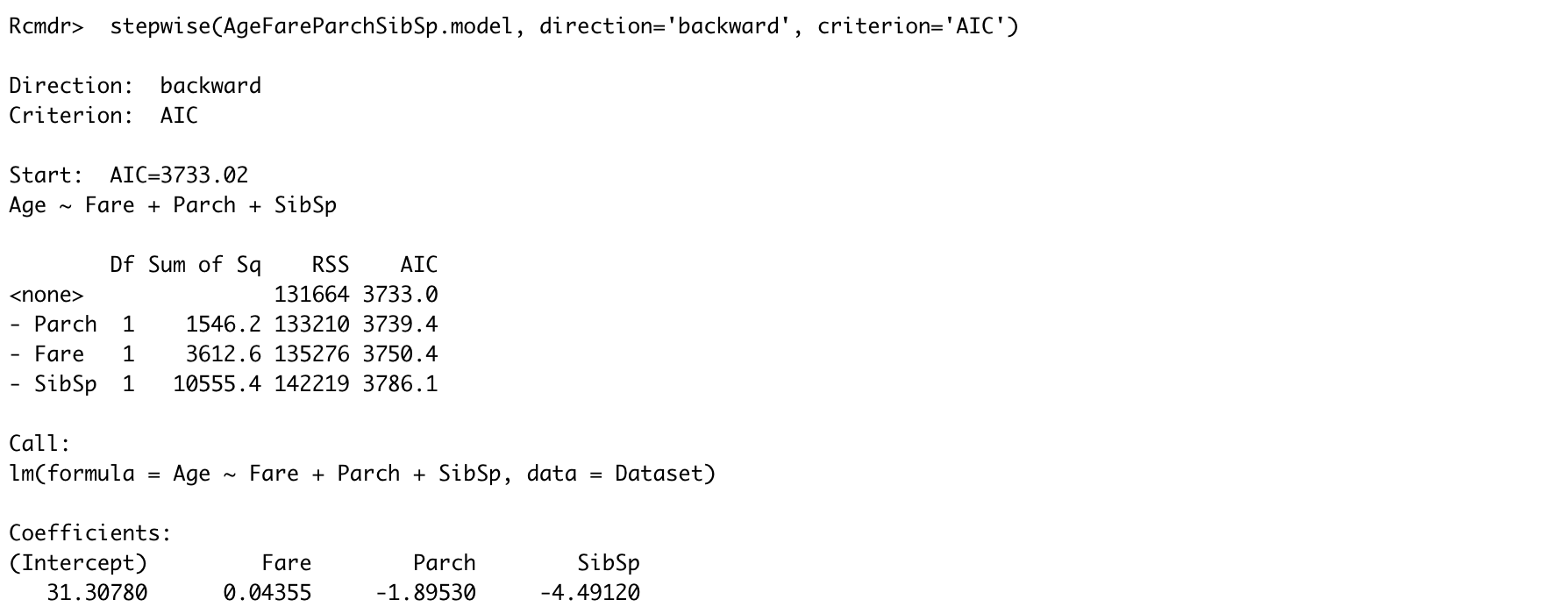
С помощью функции stepwise() производятся следующие вычисления:





Запишем уравнение оптимальной модели:

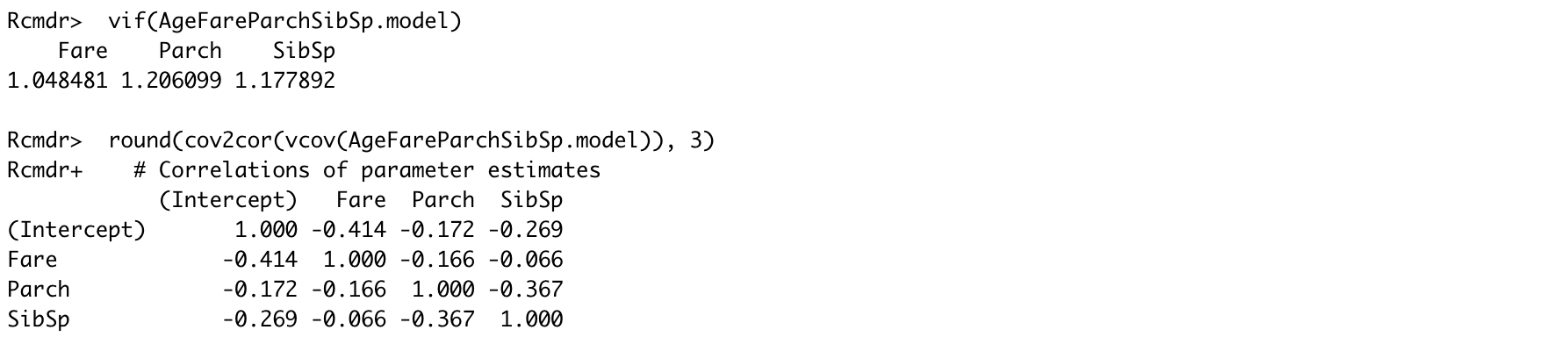
Аналогично сделаем пошаговое построение по направлению назад, заключающееся в исключении переменной. Вначале все предикаты входят в уравнение регрессии. Затем по очереди выводятся из уравнения исходя из их соответствия критерию.



Запишем уравнение оптимальной модели:

В обоих случаях коэффициенты пошаговой множественной регрессии получились такие же, как и в предыдущем пункте.

Проверим коэффициенты VIF (фактор инфляции дисперсии), характеризующий силу мультиколлинеарности, то есть чем он выше для j-го предиктора, тем сильнее линейная связь между этим и остальными предикторами.



Проанализировав данные коэффициенты, можно сделать вывод, что они невысокие, то есть значения <10.

1. Ответим на контрольные вопросы.

* Множественная регрессия.

Это регрессия, в модели которой число независимых переменных две и более. Уравнение имеет вид: *y = b0 + b1x1 + b2х2... + bkxk* .

* Пошаговая множественная регрессия.

Пошаговая вперед заключается в то, что первоначально строится модель с одной экзогенной переменной, затем добавляется следующая и строится новая модель. Модели сравниваются и, в зависимости от того ухудшилась или улучшилась модель, введенная переменная либо остается в модели, либо заменяется на другую.

Пошаговая назад начинается с того, что рассчитывается множественная регрессия на всем множестве факторов. Затем построенная модель исследуется с точки зрения статистической значимости модели в целом, статистической значимости коэффициентов регрессии, оценивается коэффициент детерминации. Затем из модели удаляется один из влияющих факторов.

* Мультиколлинеарность.

Это наличие линейной зависимости между объясняющими переменными (факторами) регрессионной модели. При этом различают полную коллинеарность, которая означает наличие функциональной (тождественной) линейной зависимости и частичную или просто мультиколлинеарность — наличие сильной корреляции между факторами.

**Выводы**

В данной лабораторной работе были исследованы возможности языка R для построения множественной линейной регрессии. Были проанализированы экспериментальные данные на мультиколлинеарность, построено уравнение множественной регрессии, а также произведена пошаговая множественная регрессия с направлением вперёд и назад.