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.

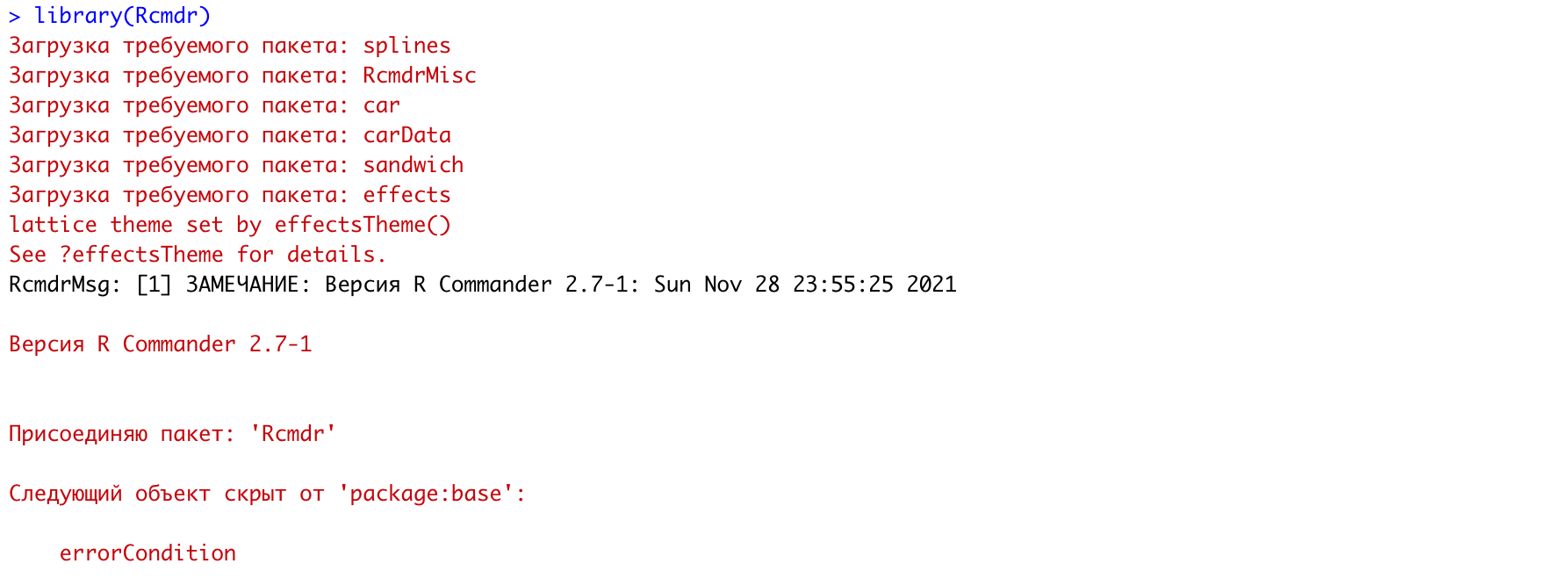
Лабораторной работа № 2.3  
«Корреляционный и регрессионный анализ данных. Исследование тесноты взаимосвязей данных в среде R»

## Цель работы

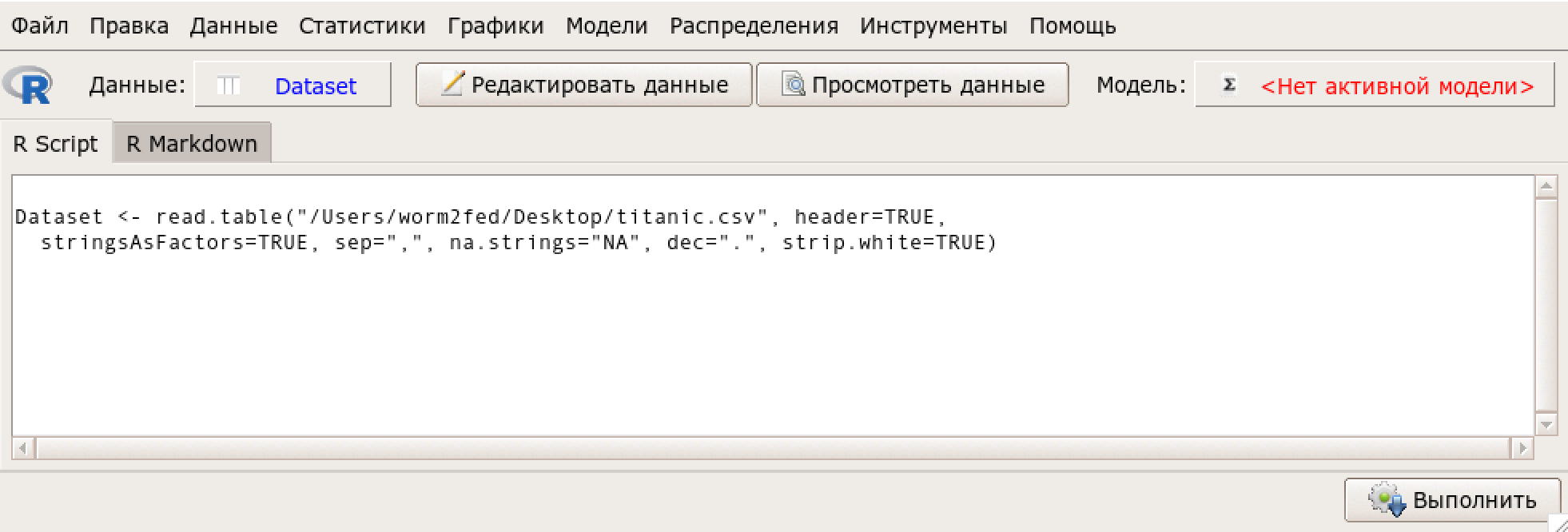
* исследовать возможности языка R для определения тесноты взаимосвязей экспериментальных данных

## Ход работы

1. Запустим пакет R commander (Rcmdr):



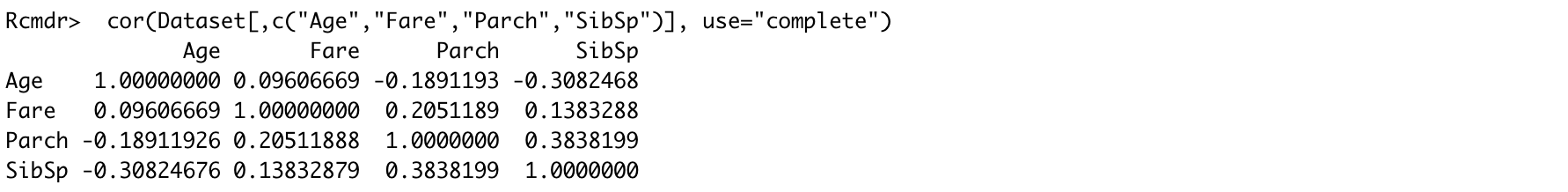
Загрузим экспериментальные данные для анализа из файла (см. рисунок 1):

  
Рисунок 1 – Импорт данных для анализа

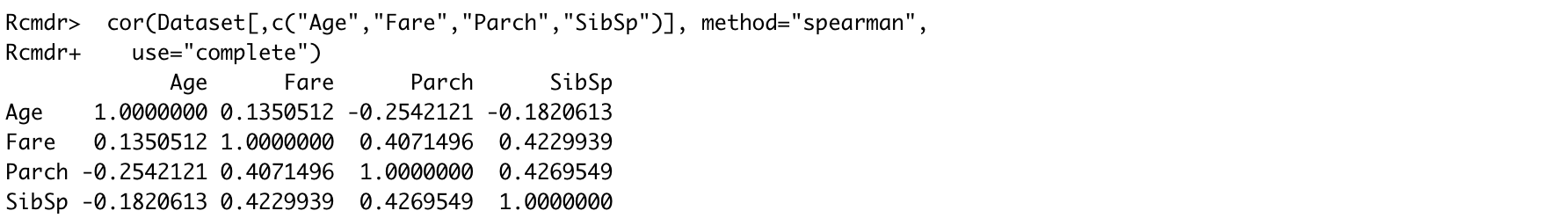
Исследуем данные при помощи корреляционных матрицы и корреляционных тестов по методам Пирсона и Спирмена.

Напомним, что Age – возраст, Fare – цена билета, Parch – количество родственников на борту 1-го порядка (мать, отец, дети), SibSp —количество родственников 2-го порядка (муж, жена, братья, сестры).

Ниже представлена корреляционная матрица по методу Пирсона:



А теперь посмотрим на матрицу по методу Спирмена:

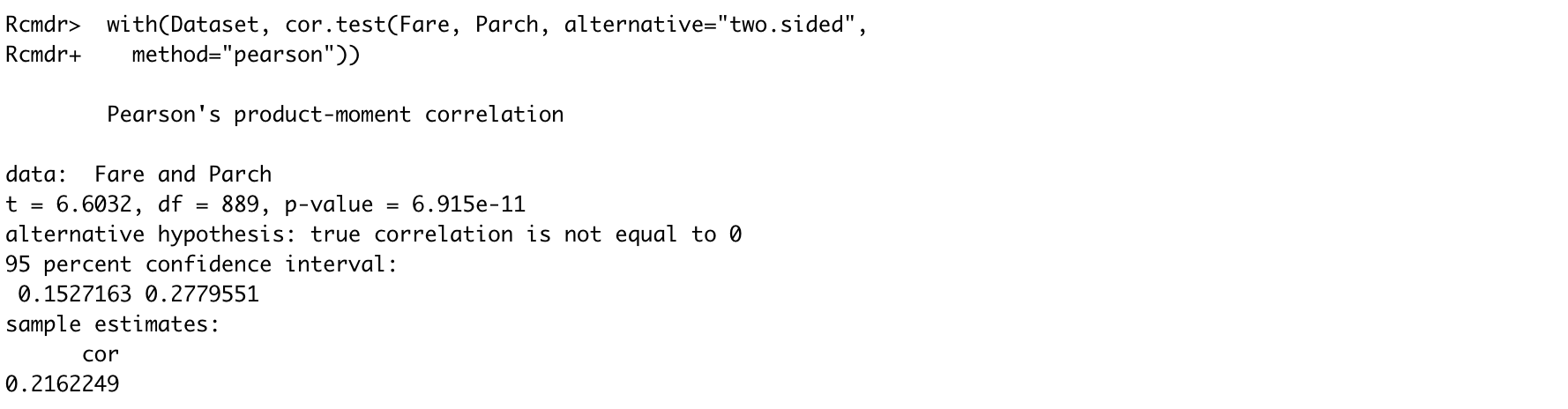


Коэффициенты корреляции в данных матрицах рассчитываются с помощью функции `cor()`. В первом случае рассчитывался коэффициент корреляции Пирсона, основанный на том, что обе анализируемые переменные распределены нормально и имеют линейную связь. А для ненормально распределенных переменных, а также при наличии нелинейной связи между ними, рассчитывался непараметрический коэффициент корреляции Спирмена. В отличие от коэффициента Пирсона, этот вариант коэффициента корреляции работает не с исходными значениями переменных, а с их рангами.

Проанализировав матрицы, можно сделать вывод, что связи:

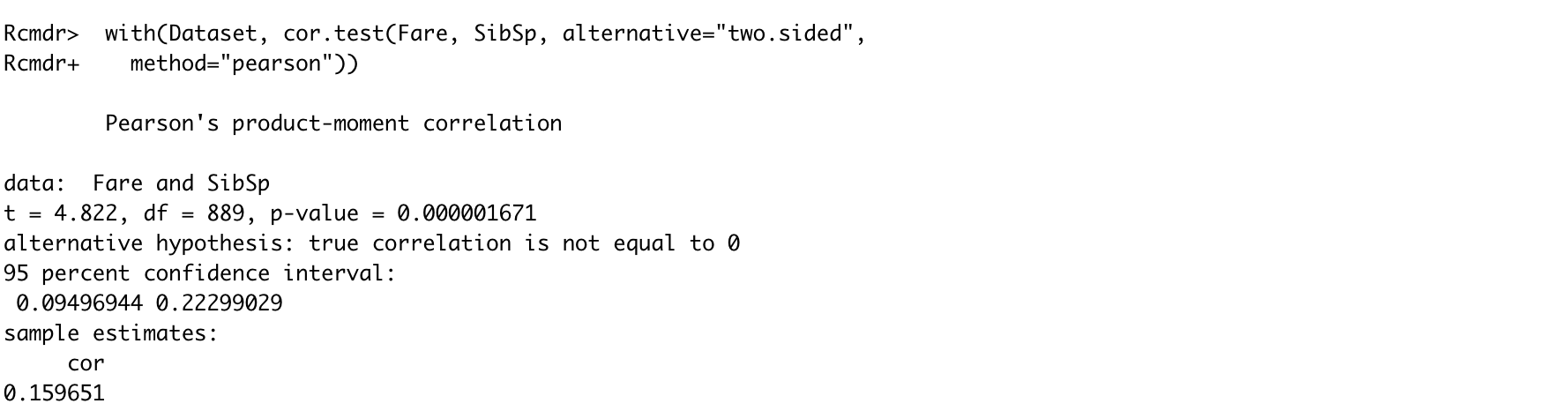
* (Age, Fare) прямая и слабая
* (Age, Parch) обратная и слабая
* (Age, SibSp) обратная и умеренная по Пирсону, но обратная слабая по Спирмену
* (Fare, Parch) и (Fare, SibSp) прямая и слабая по Пирсону, но прямая и умеренная по Спирмену
* (SibSp, Parch) прямая и умеренная.

Выполним оценку статистической значимости коэффициентов корреляции с помощью функции `cor.test()`, которая вдобавок проверяет нулевую гипотезу о равенстве коэффициента нулю:

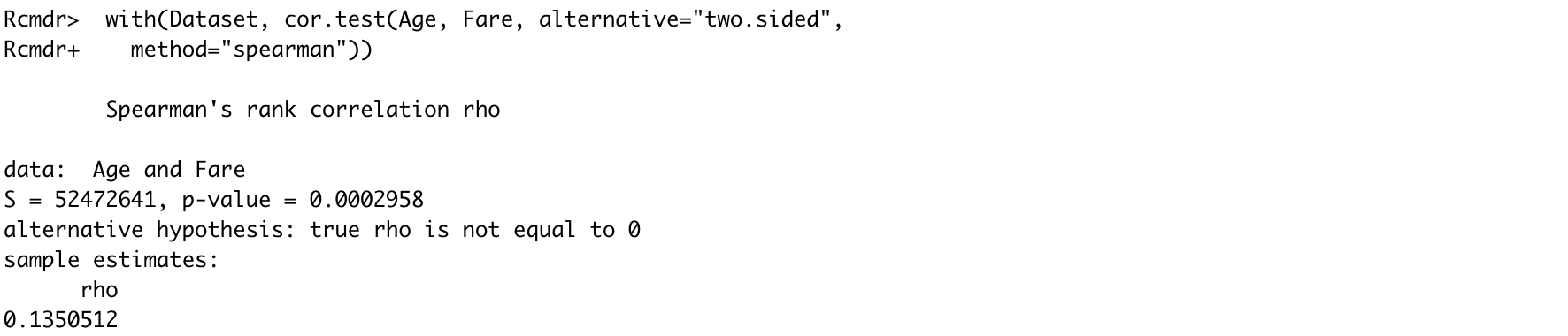


Рассчитанный коэффициент корреляции Пирсона, показывающий связь между ценой билета и количеством родственников первого порядка, оказался равен 0,2162249, что является не очень высоким показателем, но коэффициент p-value = 6,915 \* 10-11 значительно меньше, чем 0,05 – следовательно нулевая гипотеза отвергается, связь между переменными действительно существует.

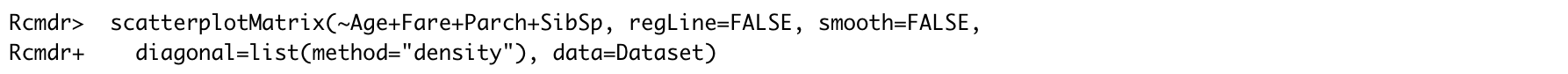
Результаты корреляционного теста для пары цена билета и количеством родственников второго порядка при коэффициенте Пирсона:



Теперь посмотрим на зависимость возраста и цены билета при коэффициенте Спирмена:

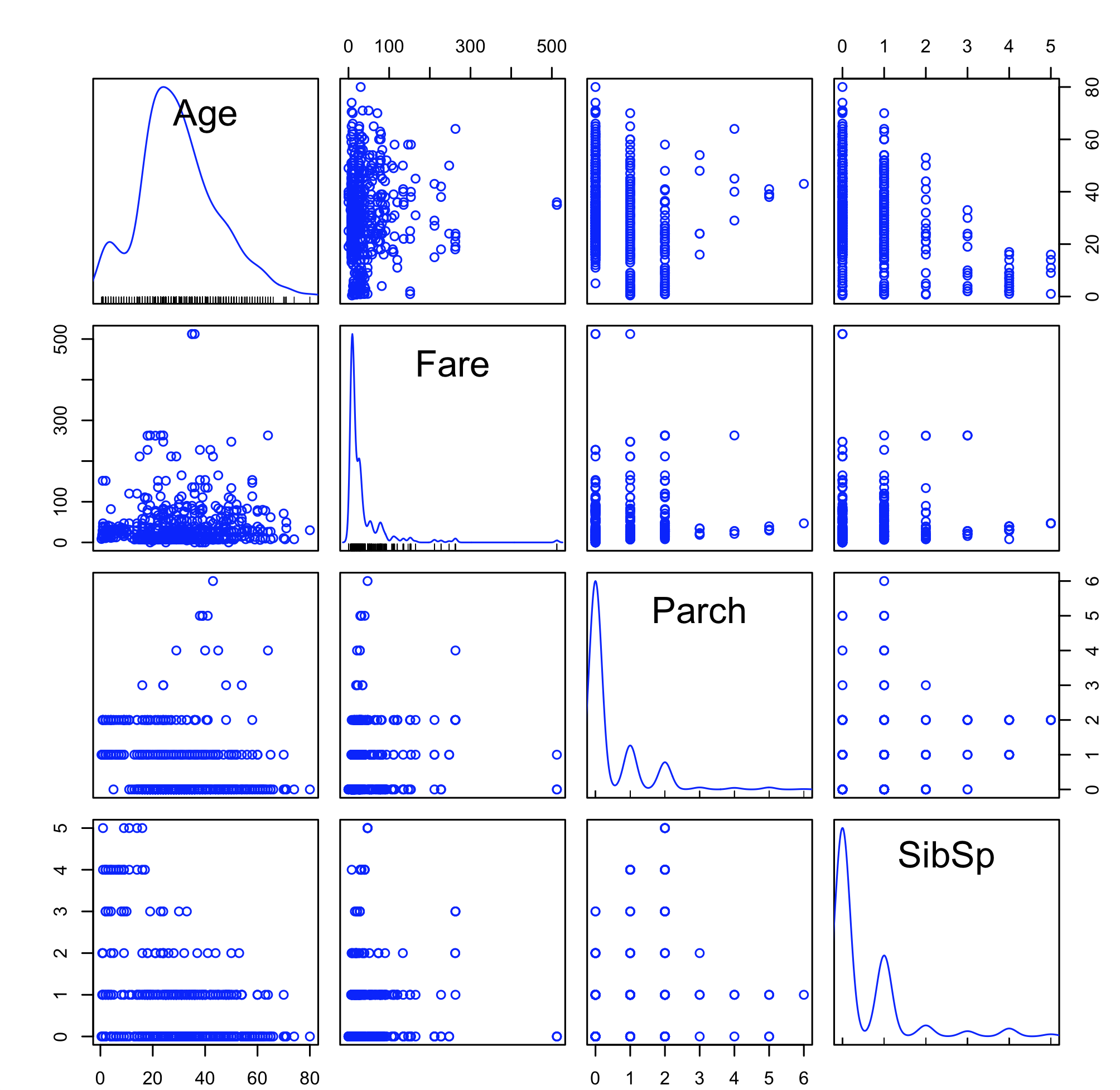


Проиллюстрируем полученные результаты при помощи матрицы точечных графиков (см. рисунок 2):

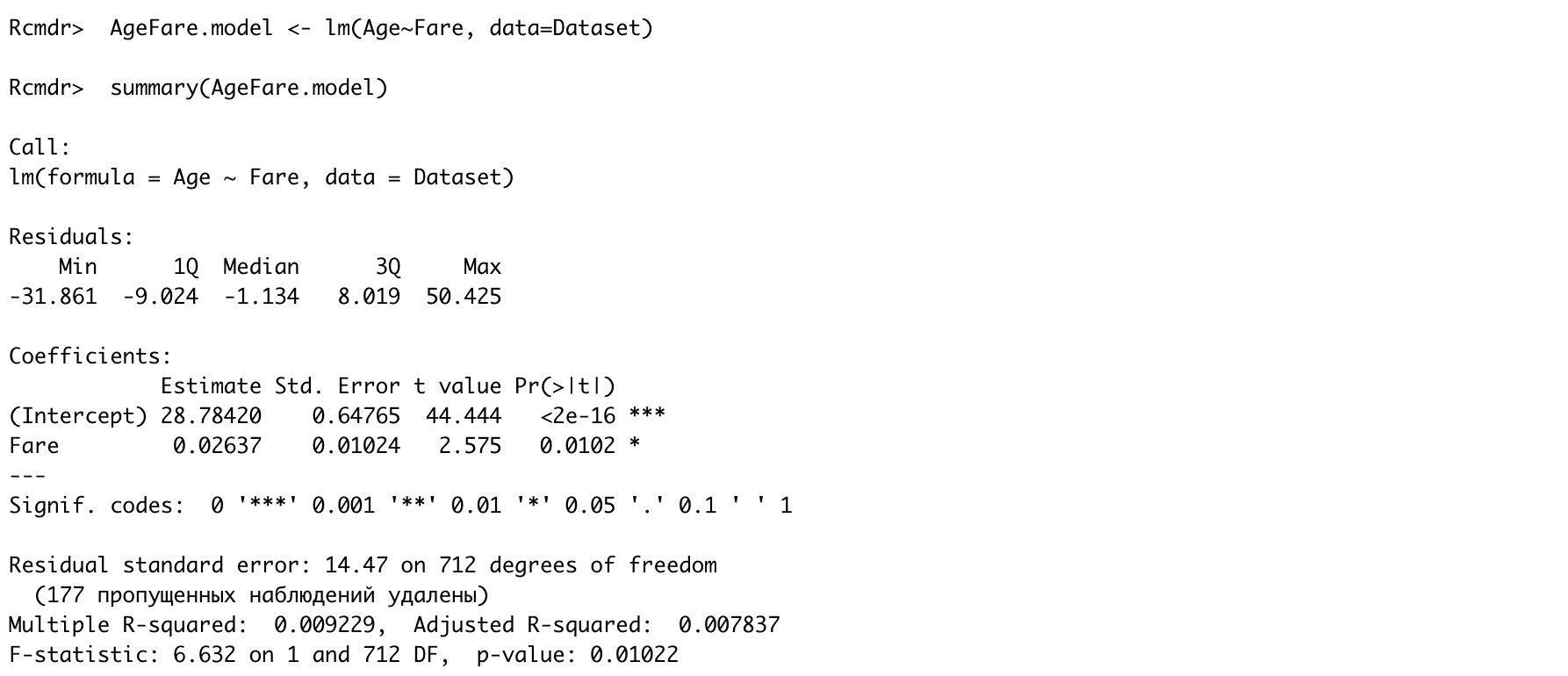


С помощью данной диаграммы можно определить потенциальные взаимосвязи между количественными переменными.

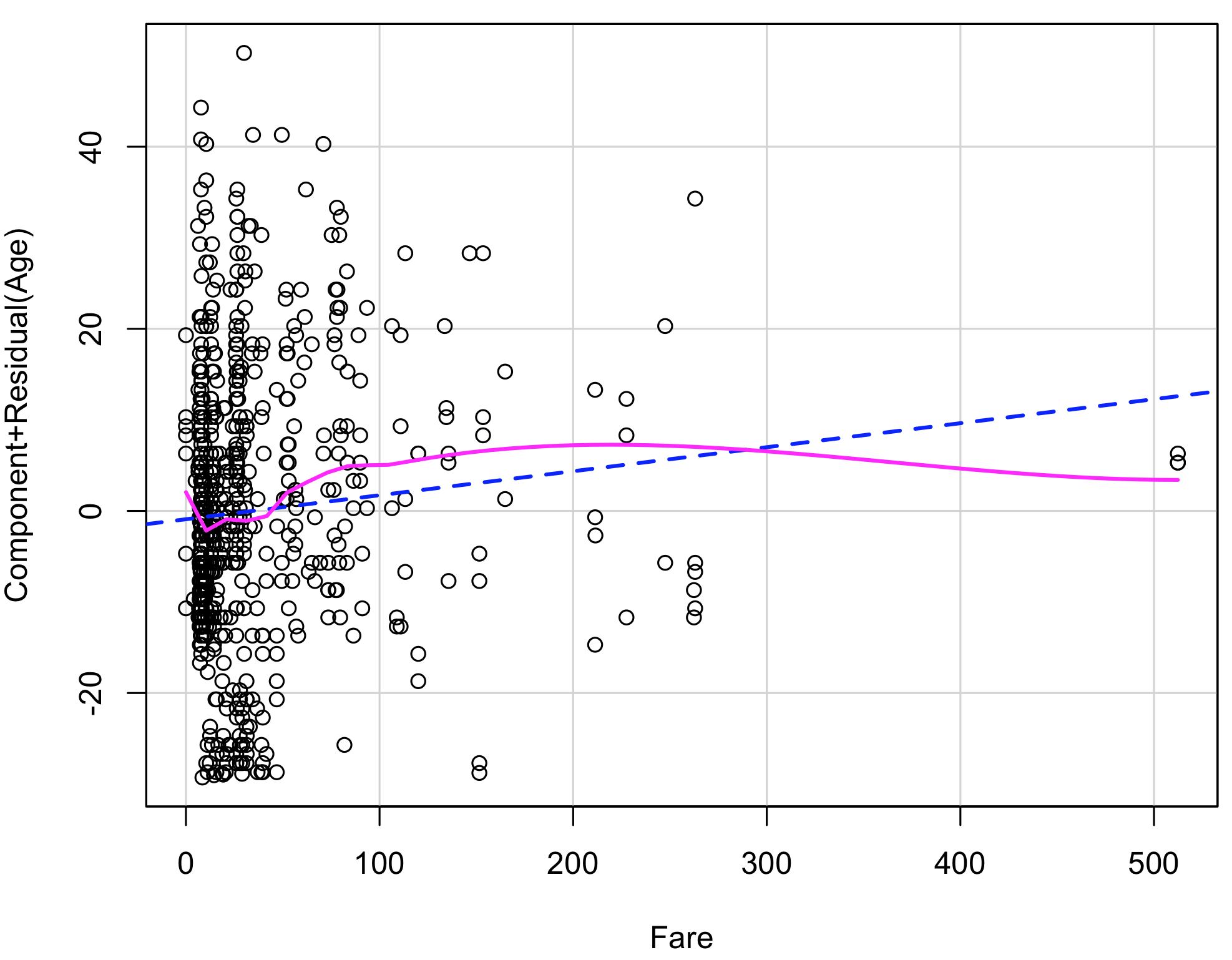
Все переменные, указанные справа от знака “~”, показываются на графике. На рисунке 2 мы имеем 4 переменные с графиком взаимосвязи для каждой пары. Матрица содержит 12 графиков: 6 в верхней панели и 6 в нижней панели от диагонали с названиями переменных.

  
Рисунок 2 – Матрица точечных графиков

Используем функцию подгонки моделей для линейной регрессии и построим график остатков для двух элементов таблицы данных (см. рисунок 3).





  
Рисунок 3 – График линейной регрессии

Из рисунка 3 видно, что аппроксимирующая и сглаживающая линии не совпадают.

Функция `lm(formula или “~”)` – основная функция в R для подгонки регрессионных моделей. Формула означает, что значение y (возраст) мы будем предсказывать по значению x (цена билета).

Составляя уравнение регрессии данной модели, опишем её в виде уравнения (1):

(1)

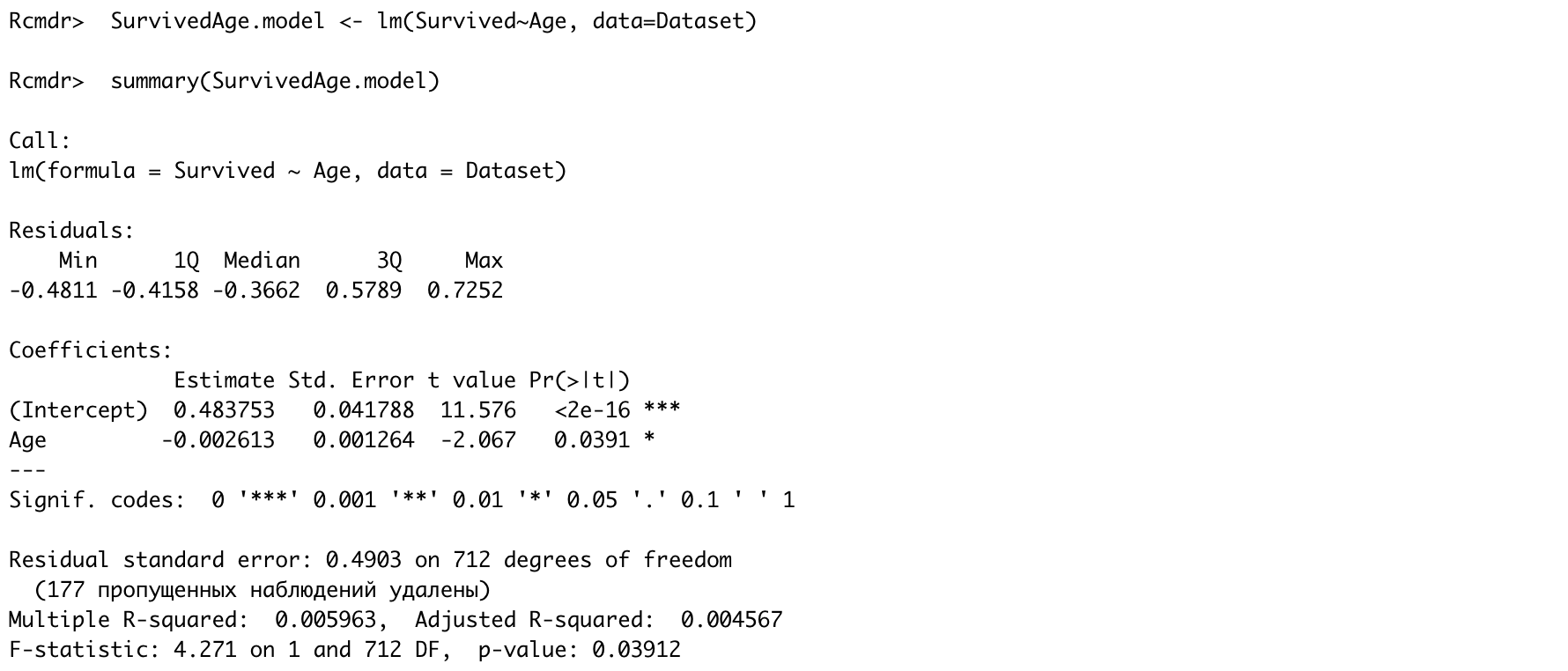
Коэффициенты k и b – минимизирующие величину ошибки, соответственно равны 0,02637 и 28,78420. Подставим коэффициенты в уравнение регрессии:

Далее проверим, насколько точно наша модель описывает данные. Для

этого используется коэффициент R-squared. Чем ближе величина этих значений к 1, тем лучше. Единица – это идеальный результат, означающий, что модель на 100% описывает данные. Но в нашей модели результат 0,009229, следовательно, только на 0.9% модель описывает данные.

Чтобы проверить, насколько предсказываемая величина зависит от предикторов (исходных данных), смотрим на значение p-value. В нашем случае оно равно 0,01022, то есть мы можем быть уверены на 98,978%, что предсказываемая величина действительно зависит от предикторов..

Произведём такой же анализ для модели, где за y (предсказываемая величина) – выжил ли человек, а за x (исходные данные) возьмём возраст. Остаточный график для модели изображён на рисунке 4.

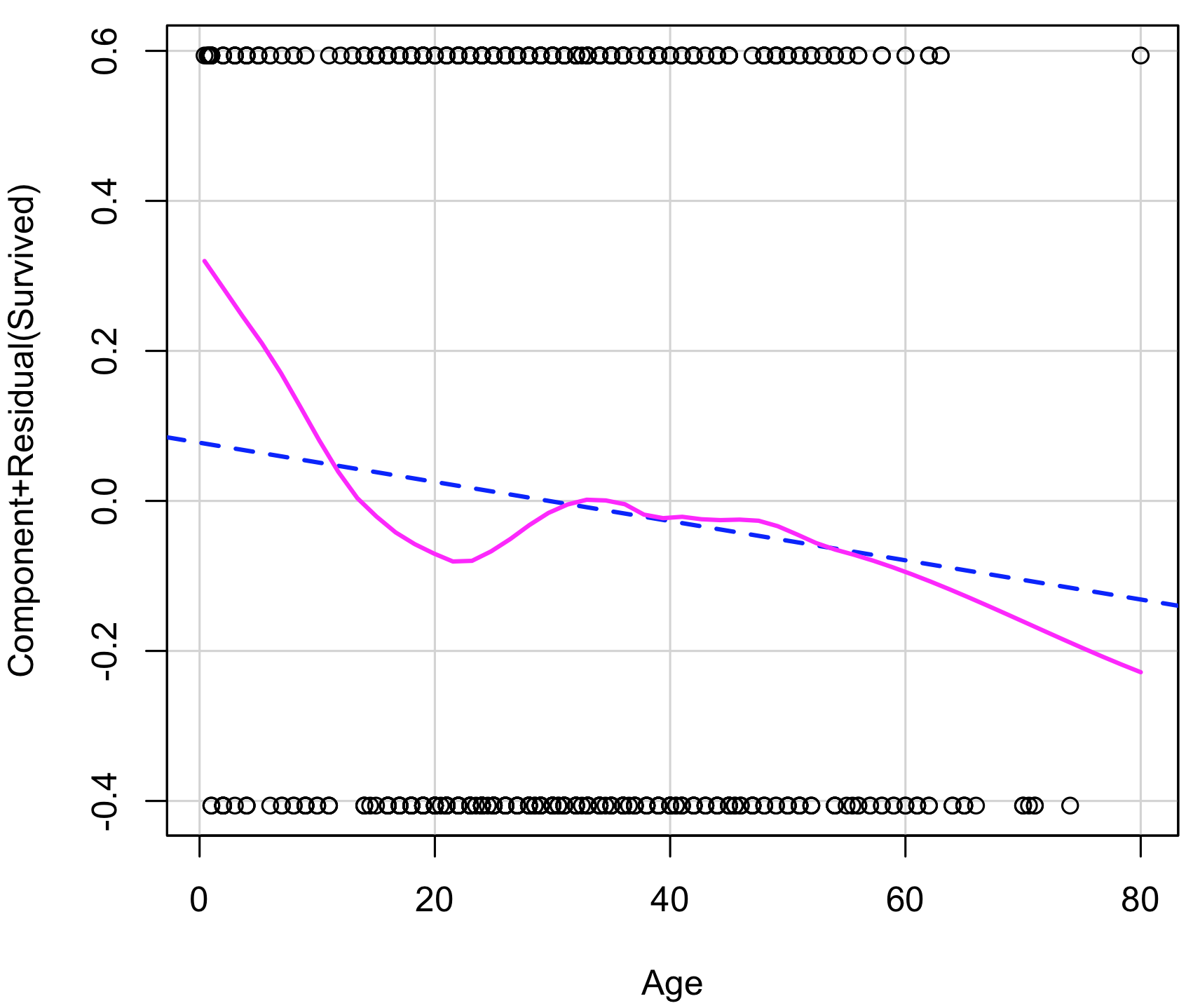




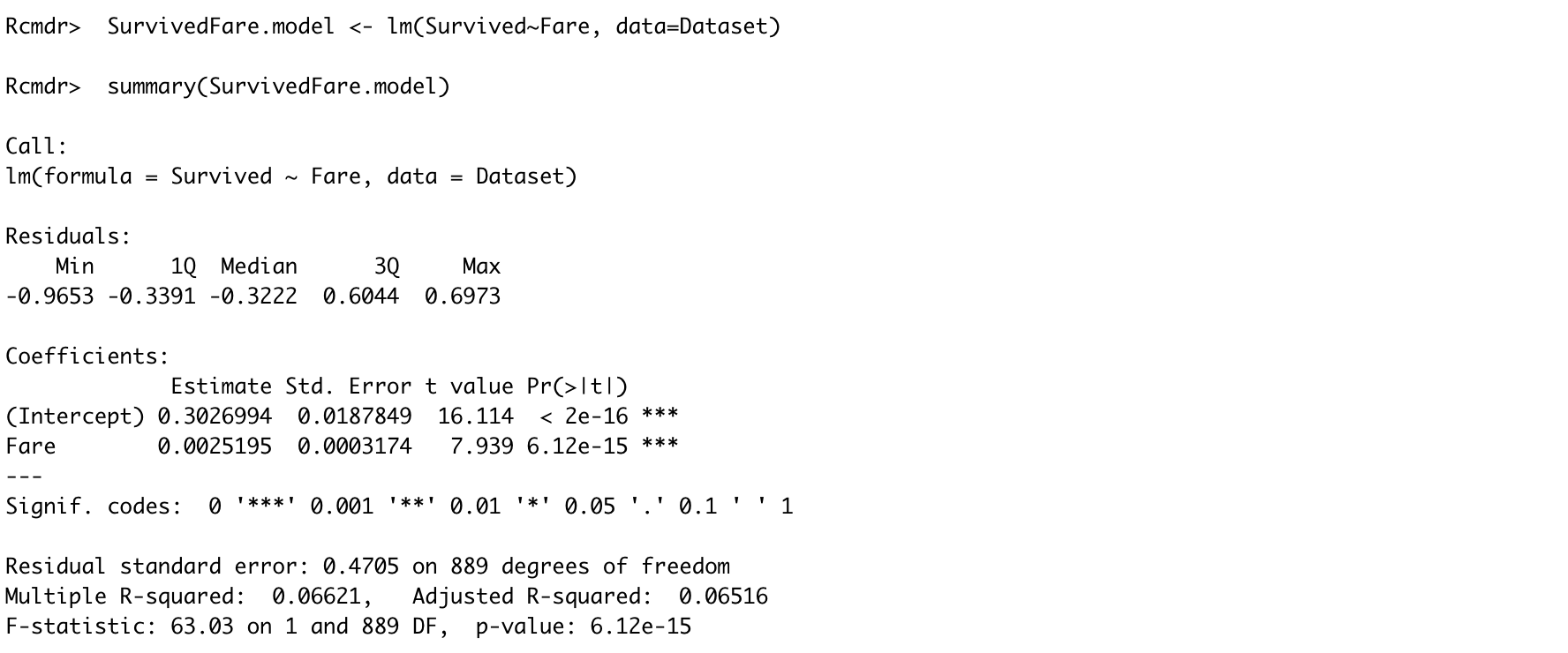
Составим уравнение регрессии данной модели:

По коэффициенту R-squared определим, что только на 0,4567% модель описывает данные.

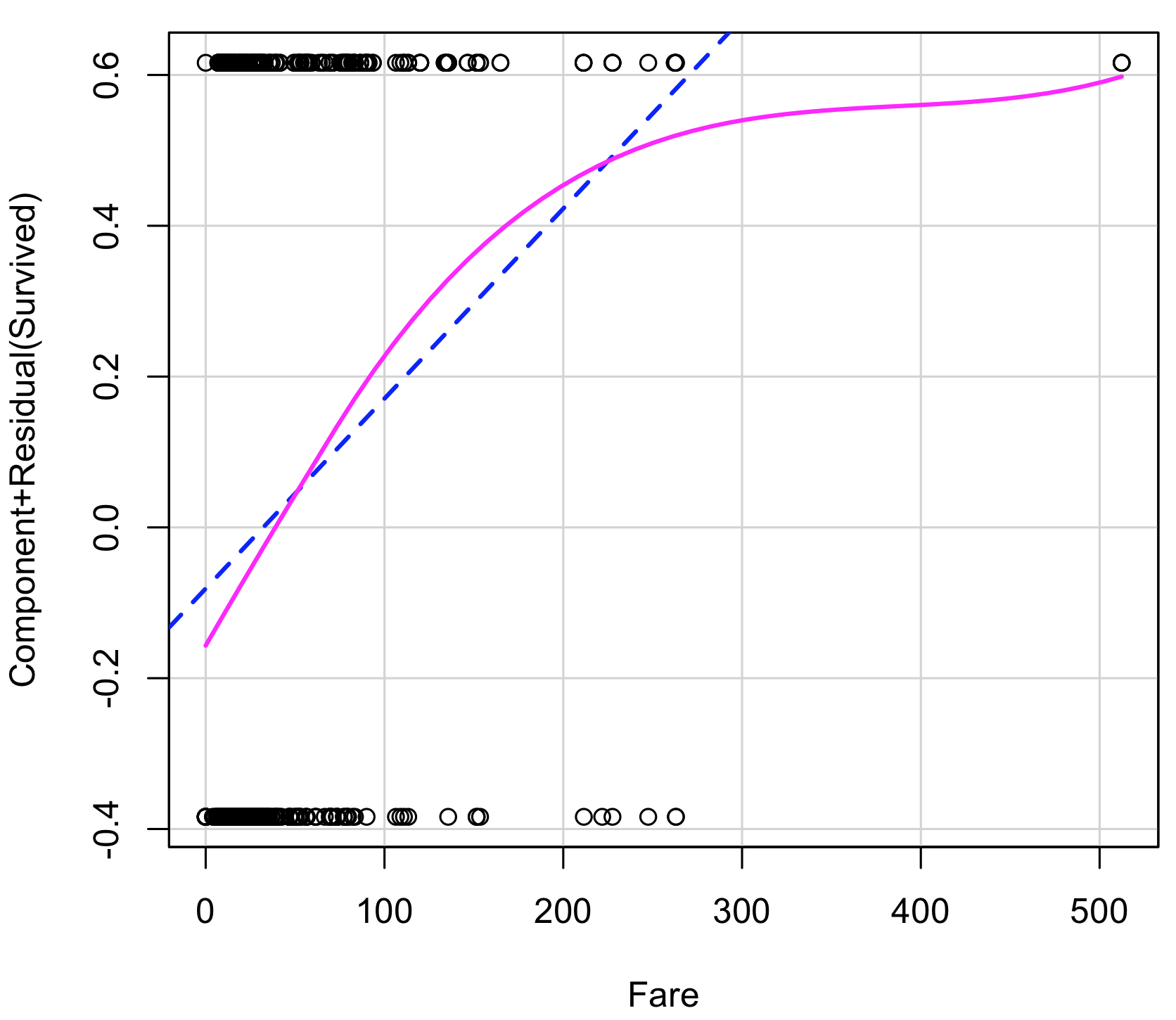
Далее смотрим на значение p-value. В нашем случае оно равно 0,03912, то есть мы можем быть уверены на 96,088%, что предсказываемая величина действительно зависит от исходных данных. Следовательно, чем p-значение меньше, тем лучше, поскольку при этом увеличивается «сила» отклонения нулевой гипотезы и увеличивается ожидаемая значимость результата.

  
Рисунок 4 – График линейной регрессии возраста и выживания

Произведём ещё один анализ для модели, где за x (исходные данные) возьмём цену билета, а за y (предсказываемая величина) – выжил ли человек. Остаточный график для модели изображён на рисунке 5.





  
Рисунок 5 – График линейной регрессии цены билета и выживания

Составим уравнение регрессии данной модели:

По коэффициенту R-squared определим, что модель описывает данные на 6,516%.

Далее смотрим на значение p-value. В нашем случае оно равно 6,12 \* 10-15, то есть предсказываемая величина значительно зависит от исходных данных.

1. Ответим на контрольные вопросы.

* Корреляционный анализ.

Устанавливает есть ли между процессами значимая связь; позволяет сделать вывод о силе взаимосвязи между парами данных.

* Регрессионный анализ.

Устанавливает, что является воздействующим фактором, а что результативным; используется для прогнозирования одной переменной на основании другой.

* Методы определения корреляционной связи.

Элементарные (параллельное сопоставлении рядов, построение корреляционной и групповой таблиц, графическое изображение с помощью поля корреляции) и статистические (расчёт коэффициентов Пирсона, Спирмена, Кендала).

* Коэффициенты корреляции.

Статистический показатель для количественной оценки существования связи между изучаемыми совокупностями случайных величин.

* Функции R для проведения корреляционного анализа.

Функция `cor()` – рассчитывает коэффициент корреляции; функция `cor.test()` – оценивает уровень значимости коэффициентов корреляции.

* Функции R для графической интерпретации корреляционного анализа.

Функция `scatterplot()` – создаёт график разброса.

* Функции R для построения линейной регрессии.

Функция `lm()` – строит линейную регрессию.

**Выводы**

В данной части лабораторной работы были исследованы возможности языка R для определения тесноты взаимосвязей экспериментальных данных. С помощью кнопочного интерфейса пакета Rcmdr были рассчитаны коэффициенты корреляции Пирсона, Спирмена, была оценена статистическая значимость коэффициентов корреляции, полученные результаты были проанализированы с помощью построенных матрицы точечных графиков и графика линейной регрессии