Министерство образования и науки РФ
ФГБОУ ВО «Кубанский государственный технологический университет»
Институт компьютерных систем и информационной безопасности
Кафедра информатики и вычислительной техники

Отчет

По лабораторной работе № 2
По дисциплине анализ и визуализация данных

Выполнил студент группы 19-КМ-ИВ1: Кирмасов Б.В.

Преподаватель: Решетняк М.Г.

Краснодар 2020

Тема: «Линейная регрессия».
Цель работы: «Изучить приемы исследования корреляционной зависимости, построения парной и множественной линейной регрессии».

Отчёт о работе.

Задание

1. **Загрузить набор данных для своего варианта, ознакомиться с его содержимым.**
2. **Построить график корреляционного поля для каждого фактора.**
3. **Построить уравнение парной линейной регрессии для каждого фактора.**
4. **Проверить значимость каждого из полученных уравнений регрессии. Показать уравнения регрессии с заданным в варианте доверительным интервалом на графиках.**
5. **Построить прогнозы по каждому из уравнений парной регрессии для заданных в варианте значений факторов.**
6. **Построить уравнение множественной линейной регрессии и получить корреляционную матрицу.**
7. **Построить прогноз по уравнению множественной регрессии для заданных в варианте значений факторов.**

1. df <- read.csv("train.csv")
2. ggplot() + geom_point(aes(x=df$Age, y=df$Fare), size = 2) + theme_bw(base_size = 18) + xlab("Возраст, лет") + ylab("Транспортные расходы") + labs(title = "Корреляционное поле")describe(df)

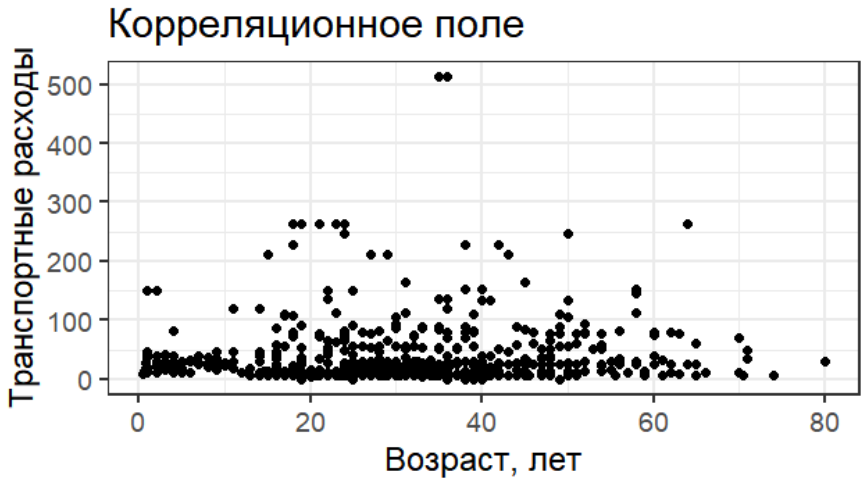

Рисунок 1 – График зависимости транспортных расходов от возраста

3. model <- lm(data=df, Fare~Age)
   model$coefficients
   Таким образом, уравнение регрессии имеет вид:
   Fare = 24.3 + 0.35*Age

Также можно посмотреть значения вектора ошибок модели — разницу между реальными расходами fare и полученной по модели. Выведем первые 10 значений этого вектора с точностью две цифры после запятой:

```
options(digits = 3)
model$residuals[1:10]
```

```
    1     2     3     4     5     7     8     9    10    11
-24.75  33.68 -25.47  16.55 -28.50   8.66  -3.93 -22.62   0.87  -9.00
```

4. summary(model)

```
Call:
lm(formula = Fare ~ Age, data = df)

Residuals:
   Min     1Q Median     3Q    Max
 -42.4  -24.5  -17.6    2.3  475.8

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   24.301      4.492    5.41  8.6e-08 ***
Age            0.350      0.136    2.58     0.01 *
---
```

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 52.7 on 712 degrees of freedom
  (177 observations deleted due to missingness)
Multiple R-squared: 0.00923, Adjusted R-squared: 0.00784
F-statistic: 6.63 on 1 and 712 DF, p-value: 0.0102

qplot(data = df, Age, Fare) + stat_smooth(method="lm", level = 0.95) +theme_bw(base_size = 18)

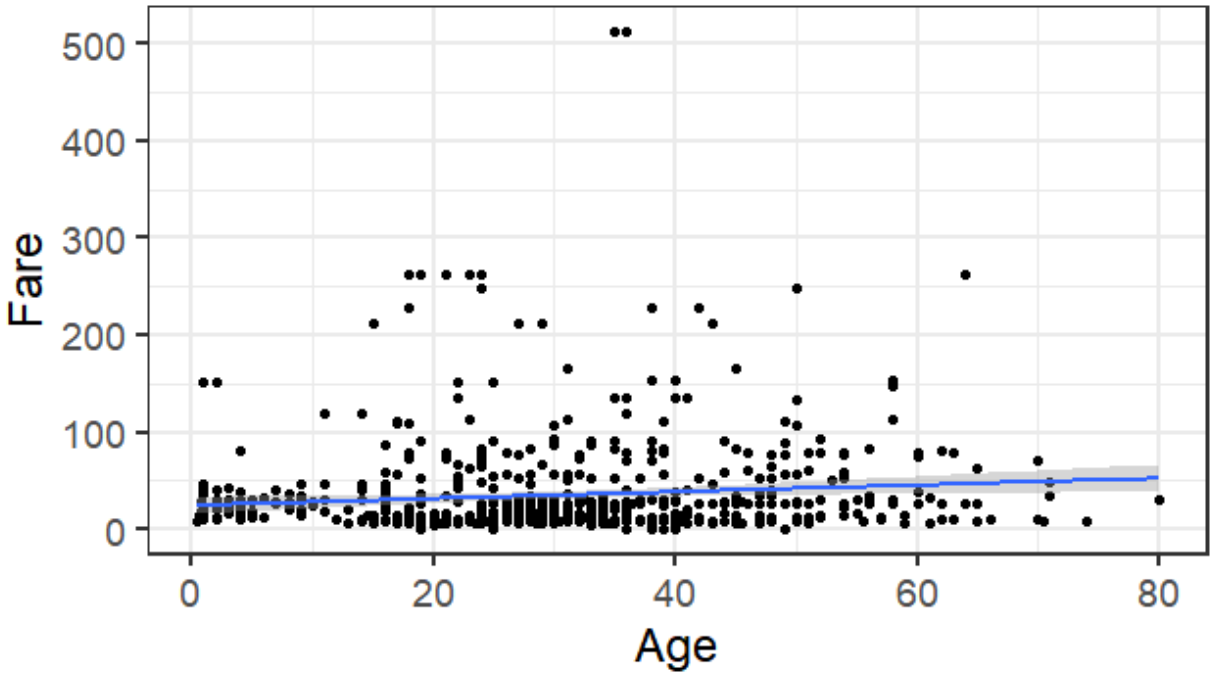

Рисунок 2 – График линейной регрессии с доверительными интервалами

5. Для того чтобы построить прогноз по полученной модели, нужно задать значения регрессора и поместить их в новый data.frame.

nd<-data.frame(Age=c(70,80,90))
predict(model,nd)

```
   1    2    3
48.8 52.3 55.8
```

6. Рассмотрим встроенный набор данных по социально-экономическим п оказателям в 47 провинциях Швейцарии в 1888 г.

Встроенный пакет graphics содержит функцию pairs, позволяющую получить все возможные диаграммы рассеяния на одном графике, а также выполнить их сглаживание с помощью опции panel.smooth:
pairs(swiss, panel= panel.smooth)

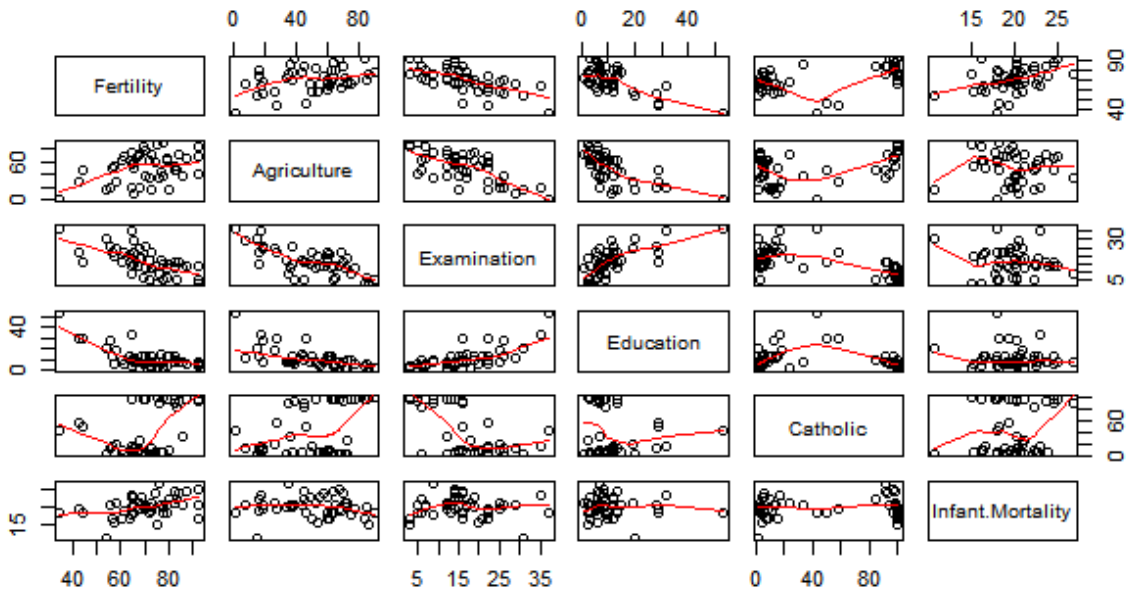

Рисунок 3 – Диаграммы рассеяния, полученные с помощью функции pairs

7. cor(swiss)

Чтобы оценить регрессию рождаемости на остальные переменные, можно воспользоваться функцией lm, а регрессоры перечислить через знак «плюс»:
model2 <- lm(data=swiss, Fertility~Agriculture+Education+Catholic)

В данном случае регрессорами стали % занятых в с/х; % католического населения и % имеющих образование выше начального.
model2$coefficients

Получить оценки коэффициентов уравнения регрессии, а также проверить основные гипотезы поможет функция summary:
summary(model2)

```
Call:
lm(formula = Fertility ~ Agriculture + Education + Catholic,
    data = swiss)

Residuals:
   Min     1Q Median     3Q    Max
-15.18  -6.55   1.38   5.82  14.84

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  86.2250     4.7347   18.21  < 2e-16 ***
Agriculture  -0.2030     0.0712   -2.85   0.0066 **
Education    -1.0721     0.1558   -6.88  1.9e-08 ***
Catholic      0.1452     0.0301    4.82  1.8e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.73 on 43 degrees of freedom
Multiple R-squared:  0.642,    Adjusted R-squared:  0.617
F-statistic: 25.7 on 3 and 43 DF,  p-value: 1.09e-09
```

Построим прогноз по аналогии с парной линейной регрессией.
Отличие заключается лишь в том, что в наборе данных необходимо указать значения каждого фактора:

```
nd2 <- data.frame(Agriculture=0.5,Catholic=0.5, Education=20)
predict(model2,nd2)
```

```
   1
64.8
```

Построение прогноза по нескольким точкам выполняется с помощью векторов значений:

```
nd2 <- data.frame(Agriculture=c(0.5,0.8),Catholic=c(0.5, 0.65),Education=c(20, 25))
predict(model2,nd2)
```

```
   1    2
64.8 59.4
```

Вывод: Мы изучили приемы исследования корреляционной зависимости, по строения парной и множественной линейной регрессии.