**Задание. Линейная регрессия – цена автомобиля – включение категориальной переменной**

Для выполнения задания были загружены данные из файла

“kuiper.tsv” и пропущены первые четыре строки поскольку они не содержат данных.

install.packages(c("rstudioapi"), repos = "https://cran.r-project.org/")

setwd(dirname(rstudioapi::getActiveDocumentContext()$path))

df <- read.delim("kuiper.tsv", stringsAsFactors = TRUE, skip = 4)

Далее проверяем, что данные прочитаны корреткно

head(df)

В результате получаем

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| row | Price | Mileage | Make | Model | Trim | Type | Cylinder | Liter | Doors | Cruise | Sound | Leather |
| 1 | 17314.1 | 8221 | Buick | Century | Sedan 4D | Sedan | 6 | 3.1 | 4 | 1 | 1 | 1 |
| 2 | 17542.04 | 9135 | Buick | Century | Sedan 4D | Sedan | 6 | 3.1 | 4 | 1 | 1 | 0 |
| 3 | 16218.85 | 13196 | Buick | Century | Sedan 4D | Sedan | 6 | 3.1 | 4 | 1 | 1 | 0 |
| 4 | 16336.91 | 16342 | Buick | Century | Sedan 4D | Sedan | 6 | 3.1 | 4 | 1 | 0 | 0 |
| 5 | 16339.17 | 19832 | Buick | Century | Sedan 4D | Sedan | 6 | 3.1 | 4 | 1 | 0 | 1 |
| 6 | 15709.05 | 22236 | Buick | Century | Sedan 4D | Sedan | 6 | 3.1 | 4 | 1 | 1 | 0 |

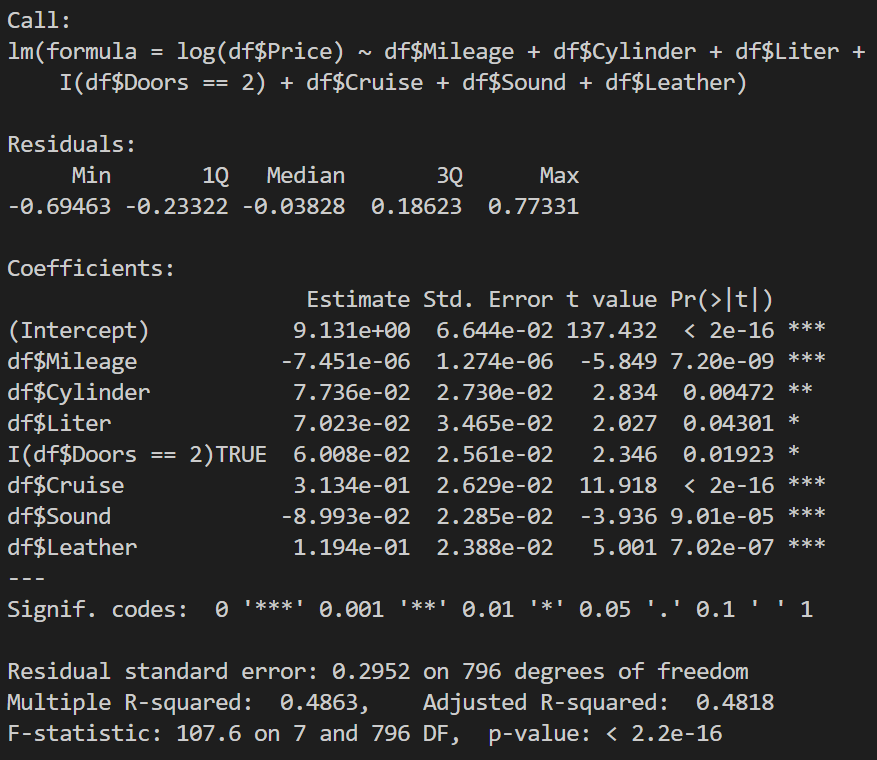
Сравнив с оригиналом, убедились, что данные загружены верно.

Далее строим регрессию логарифма цены от всех количественных и индикаторных переменных. Вместо количества дверей берём индикаторную переменную для 2-ух дверей

reg <- lm(log(df$Price) ~ df$Mileage + df$Cylinder + df$Liter +

I(df$Doors == 2) + df$Cruise + df$Sound + df$Leather)

summary(reg)



По результатам оценки регрессии можно сказать, что все переменные значимыми на уровне значимости 0,05, а значение p-value низкое, это значит регрессия хорошо описывает наши данные.

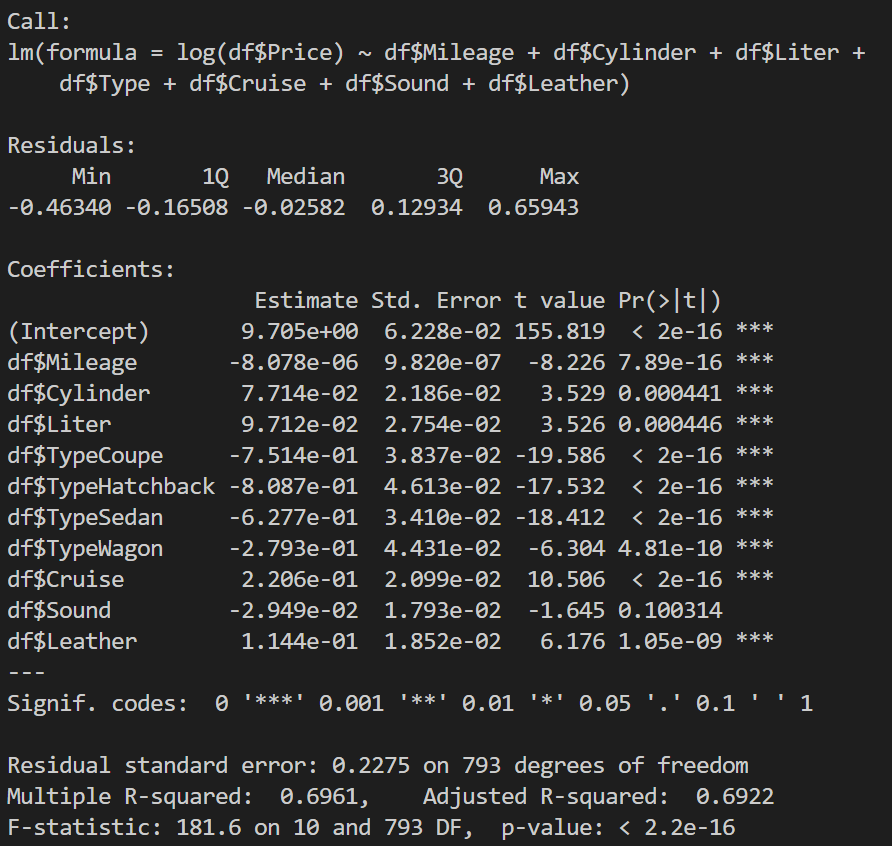
Далее вместо количества дверей возмём категориальную переменную типа кузова и сравнимм показатели точности подгонки регрессии.

new\_reg <- lm(log(df$Price) ~ df$Mileage + df$Cylinder + df$Liter + df$Type +

df$Cruise + df$Sound + df$Leather)

summary(new\_reg)

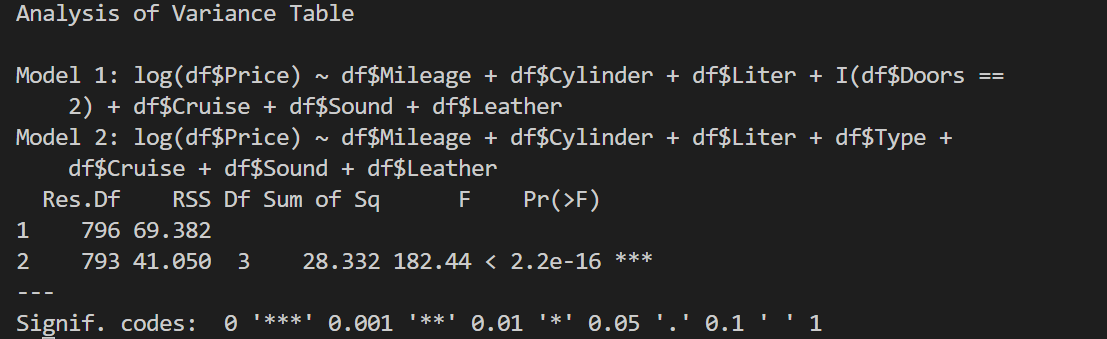
table(df$Type, df$Doors)



|  |  |  |
| --- | --- | --- |
|  | 2 | 4 |
| Convertible | 50 | 0 |
| Coupe | 140 | 0 |
| Hatchback | 0 | 60 |
| Sedan | 0 | 490 |
| Wagon | 0 | 64 |

Проведём тест на добавление переменной

anova(reg, new\_reg) # Проводим тест на добовление пременной

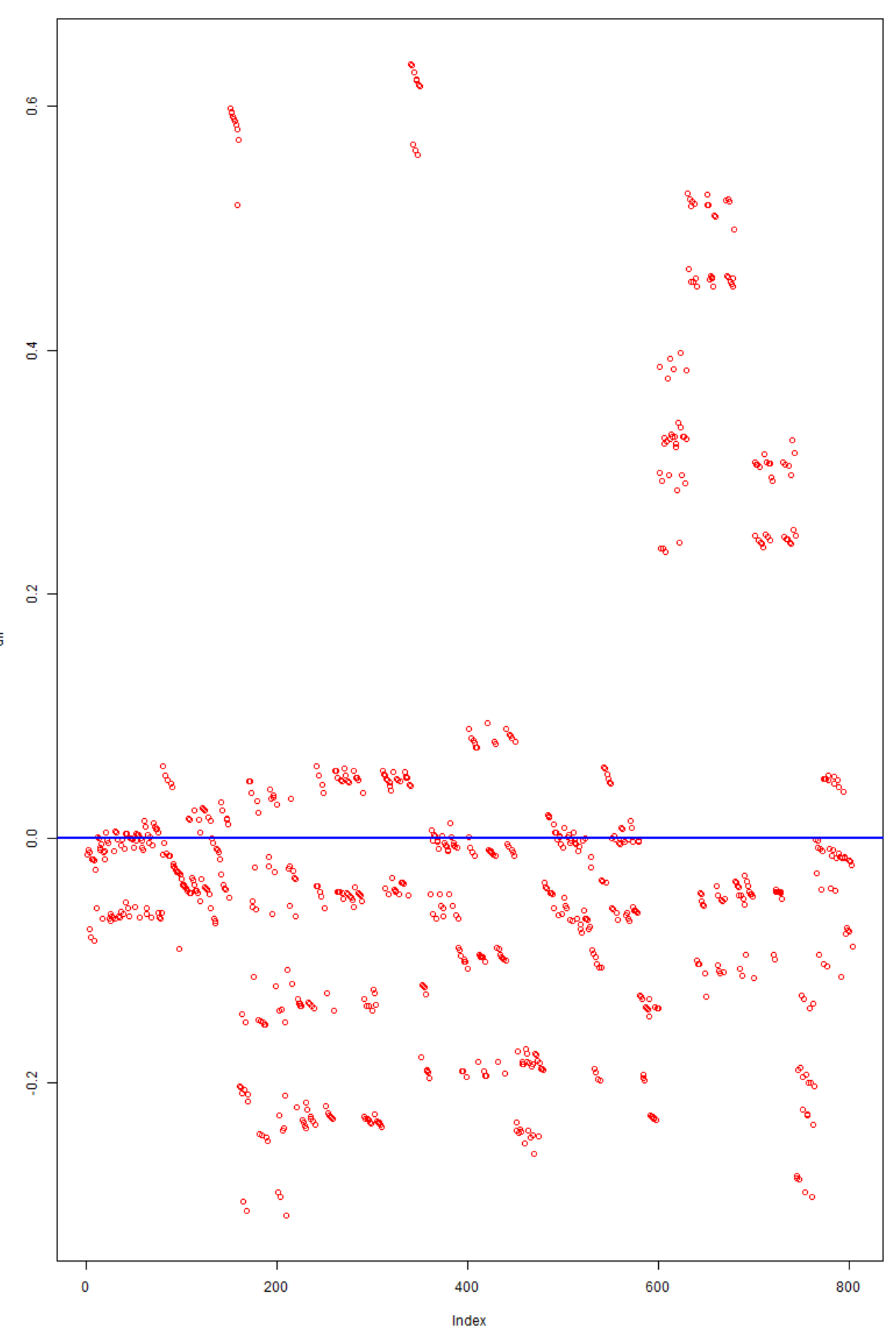


Найдём разность расчётных значений для двух регрессий

dif <- fitted(new\_reg) - fitted(reg)

plot(dif, asp = 1) # Анализируем различие

abline(h = 0, col = "#0000ff", lwd = 2) # Добавляем линию нуля



Задание. Линейная регрессия – прогноз интенсивности дорожного

движения в США

Для выполнения задания были загружены данные из файла

“kuiper.tsv” и пропущены первые пять строки поскольку они не содержат данных.

setwd(dirname(rstudioapi::getActiveDocumentContext()$path))

df <- read.delim("StreetLight2020.tsv", stringsAsFactors = TRUE, skip = 5)

Далее проверяем, что данные прочитаны корреткно

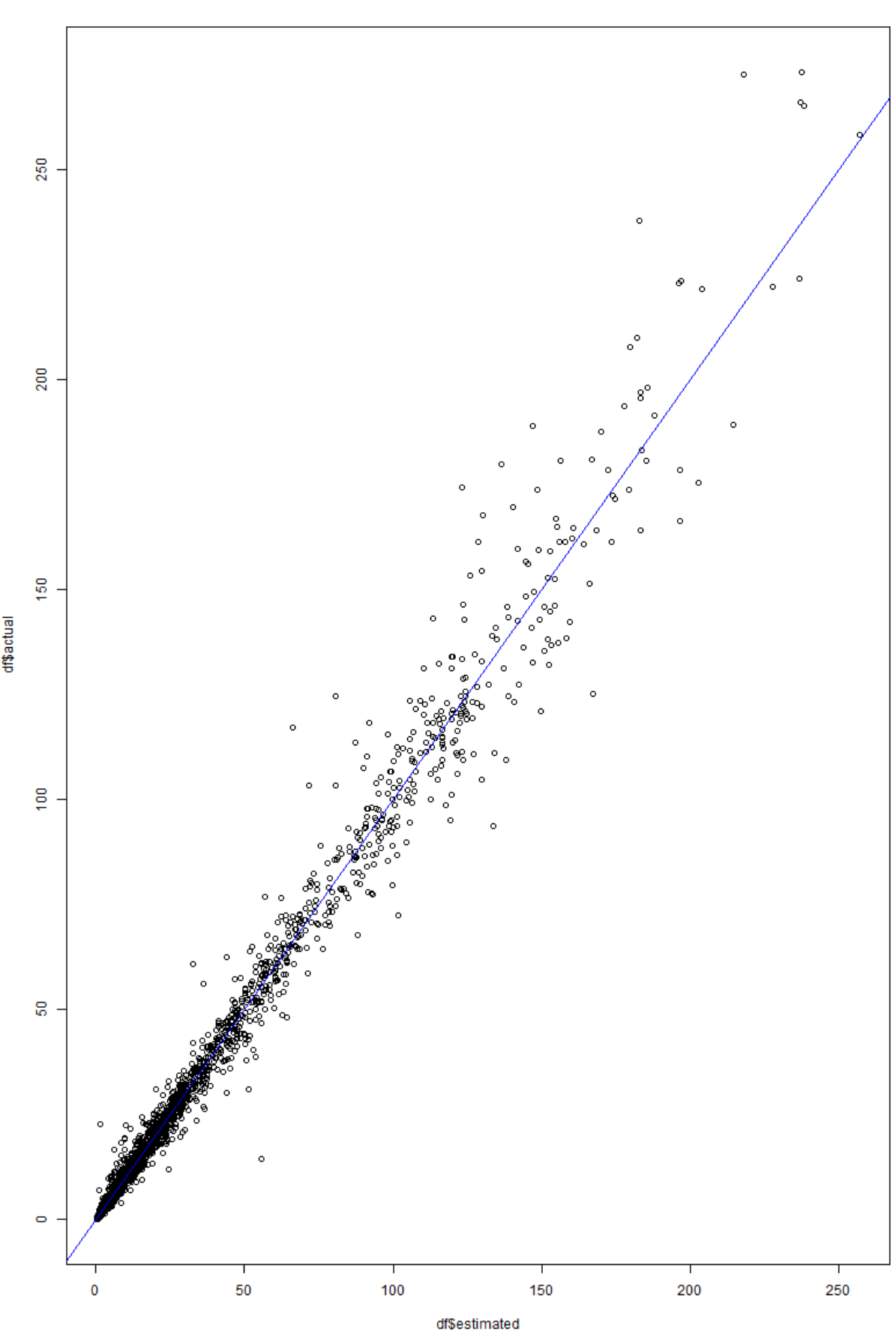
head(df)

|  |  |  |
| --- | --- | --- |
|  | actual | estimated |
| 1 | 272.809 | 217.935 |
| 2 | 178.372 | 196.553 |
| 3 | 27.0767 | 21.6445 |
| 4 | 29.8311 | 27.2528 |
| 5 | 164.797 | 155.192 |
| 6 | 99.6748 | 104.367 |

Построим точечную диаграмму фактических actual от estimated и добавим на данную диаграмму “теоретически правильную” линию регрессии.

plot(df$actual ~ df$estimated)

abline(c(0, 1), col = "#0000ff", lwd = 1)

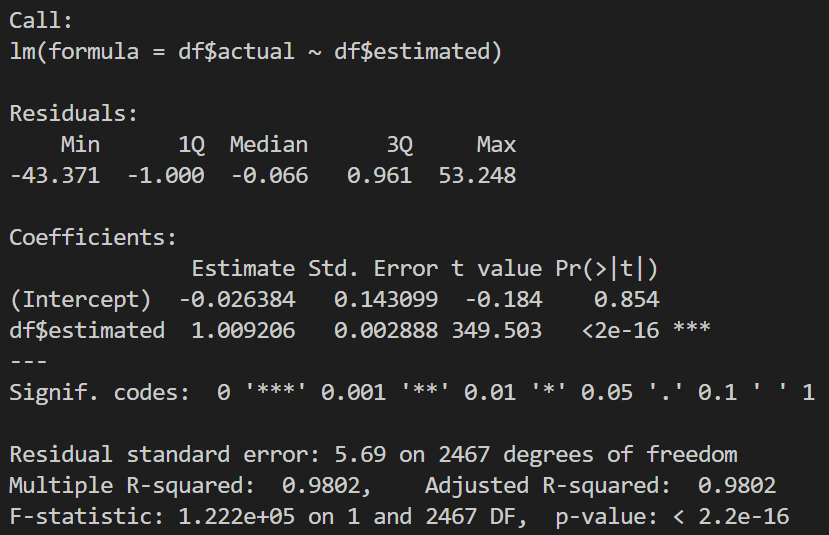


Далее строим регрессию actual от estimated и смотрим основные статистические параметры.

reg <- lm(df$actual ~ df$estimated)

summary(reg)

confint(reg)



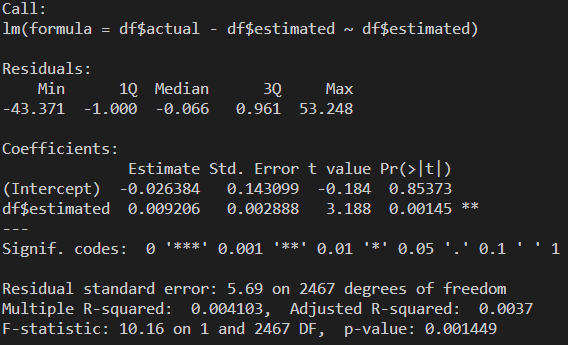
|  |  |  |
| --- | --- | --- |
|  | 2.5 | 97.5 |
| (Intercept) | -0.30699 | 0.254223 |
| df$estimated | 1.003544 | 1.014868 |

Проведём тест на равенство коэффициента при estimate равном нулю

reg\_1 <- lm(df$actual - df$estimated ~ df$estimated)

summary(reg\_1)

confint(reg\_1)



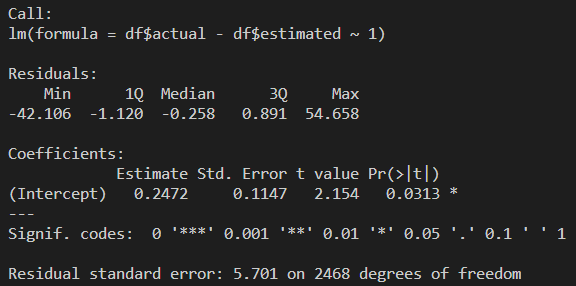


Ноль не входит в границы доверительного интервала, поэтому коэффициент не равен 0.

Строим регрессию ошибок от константы

reg\_2 <- lm(df$actual - df$estimated ~ 1)

summary(reg\_2)

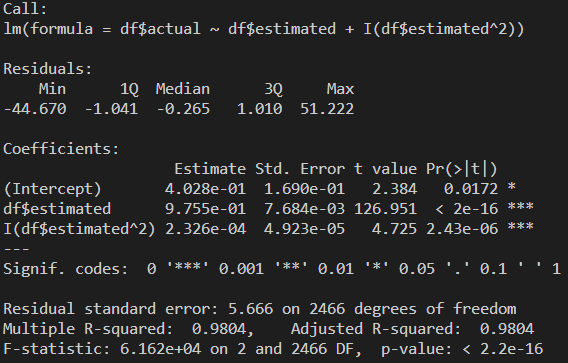


По результатам видим, что константа не равна 0, а p-value равно 0.0313.

Добавляем в регрессию фактических от прогнозных значений прогнозные значения в квадрате

reg\_3 <- lm(df$actual ~ df$estimated + I(df$estimated^2))

summary(reg\_3)



Нулевая гипотеза отвергается поскольку все коэффициенты оказались значимыми на уровне значимости 0.05, значение p-value низкое, а значит модель хорошо описывает наши данные, и переменная estimate в квадрате является значимой на уровне значимости 0.001.

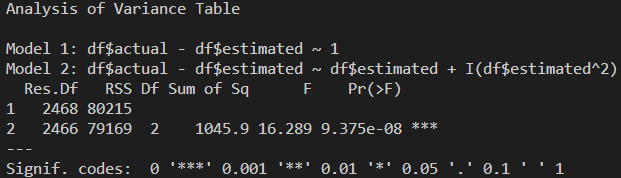
Проведём тесты с помощью anova()

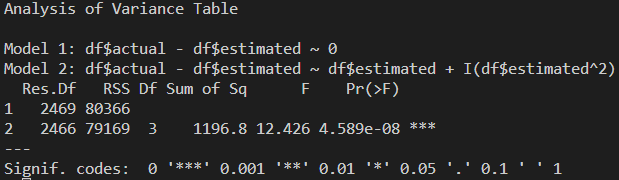
reg\_5 <- lm(df$actual - df$estimated ~ 1)

anova(reg\_5, reg\_4)

reg\_5 <- lm(df$actual - df$estimated ~ 0)

anova(reg\_5, reg\_4)





Различия в RSS значимы, а значит переменные значимы одновременно и по результатам тестов можно сказать, что коэффициенты не равны нулю.