**Задание МЛР1 – нахождение параметров множественной линейной регрессии. Прогнозирование.**

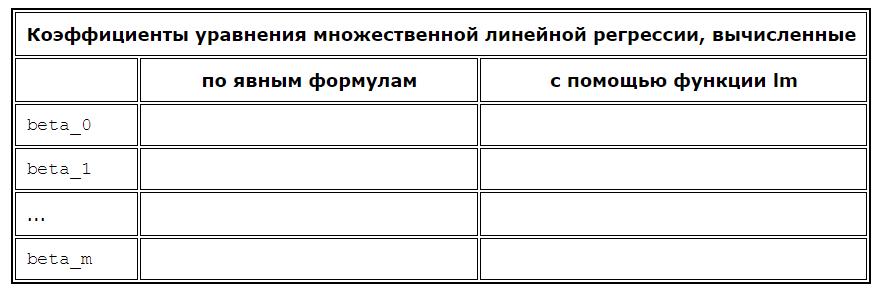
Выполнил студент 2 курса

группы 09-715(1)

Санамян Артак Размикович

**Текст задания:**

Решите в пакете R задачу построения и анализа уравнения множественной линейной регрессии:

1. Подберите данные для задачи;
2. Постройте с помощью пакета R уравнение линейной регрессии (сделайте это 2мя способами:
   * + по явным формулам;
     + с помощью функции lm.
3. Сравните результаты с помощью таблицы следующего вида:   
     
   
4. Приведите в отчёте матрицу попарных корреляций и сводную информацию (summary) модели.
5. Проанализируйте степень влияния факторов на переменную отклика согласно каждой из полученных моделей;
6. Сделайте прогноз (с помощью каждой из полученных моделей. если они различны) - на вход модели подайте не менее 10 значений из обучающей выборки и сравните:
   * + истинное значение переменной отклика;
     + прогноз, полученный с помощью модели, коэффициенты которой вычислены явно;
     + прогноз, полученный с помощью модели, коэффициенты которой вычислены с помощью функции lm.
7. Проанализируйте и объясните полученные результаты.
8. Опишите полученные результаты в файле формата Word и прикрепите файл здесь.

Источник числовых данных, использованный для выполнения задания: [http://archive.ics.uci.edu/ml/datasets/Auto+MPG](http://archive.ics.uci.edu/ml/datasets/Auto+MPG%09)

**Описание выполненной работы:**

Исследуем зависимость расхода топлива в городском цикле в милях на галлон (mpg) от следующих факторов:

* + - число цилиндров;
    - смещение;
    - мощность;
    - вес;
    - ускорение.

Перед применением методов проведем нормализацию данных по формуле:

Рассмотрим сводную информацию (summary) модели.

Residuals:

Min 1Q Median 3Q Max

-0.30802 -0.07611 -0.00905 0.05968 0.43462

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.6750986 0.0348099 19.394 <2e-16 \*\*\*

cylinders -0.0529160 0.0545937 -0.969 0.3330

displacement -0.0008556 0.0933718 -0.009 0.9927

horsepower -0.2214708 0.0815369 -2.716 0.0069 \*\*

weight -0.4865494 0.0766040 -6.351 6e-10 \*\*\*

acceleration -0.0130042 0.0561912 -0.231 0.8171

---

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

Residual standard error: 0.113 on 386 degrees of freedom

Multiple R-squared: 0.7077, Adjusted R-squared: 0.7039

F-statistic: 186.9 on 5 and 386 DF, p-value: < 2.2e-16

Коэффициент детерминации равен 0.7077, исправленный коэффициент детерминации равен 0.7039.

Вычислим матрицу попарных корреляций между предикторами:

cylinders displacement horsepower weight acceleration mpg

cylinders 1.0000000 0.9508233 0.8429834 0.8975273 -0.5046834 -0.7776175

displacement 0.9508233 1.0000000 0.8972570 0.9329944 -0.5438005 -0.8051269

horsepower 0.8429834 0.8972570 1.0000000 0.8645377 -0.6891955 -0.7784268

weight 0.8975273 0.9329944 0.8645377 1.0000000 -0.4168392 -0.8322442

acceleration -0.5046834 -0.5438005 -0.6891955 -0.4168392 1.0000000 0.4233285

mpg -0.7776175 -0.8051269 -0.7784268 -0.8322442 0.4233285 1.0000000

Заметим, что имеется сильная положительная корреляция между числом цилиндров (переменная cylinders) и смещением (displacement), а также между весом (weight) и смещением (displacement). Поэтому исключим факторы cylinders и displacement.

Пересчитаем корреляционную матрицу:

horsepower weight acceleration mpg

horsepower 1.0000000 0.8645377 -0.6891955 -0.7784268

weight 0.8645377 1.0000000 -0.4168392 -0.8322442

acceleration -0.6891955 -0.4168392 1.0000000 0.4233285

mpg -0.7784268 -0.8322442 0.4233285 1.0000000

Получим новую регрессионную модель:

Residuals:

Min 1Q Median 3Q Max

-0.29466 -0.07276 -0.00880 0.05771 0.43249

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.668582 0.034120 19.595 < 2e-16 \*\*\*

horsepower -0.232425 0.078245 -2.970 0.00316 \*\*

weight -0.543064 0.054178 -10.024 < 2e-16 \*\*\*

acceleration -0.000923 0.055108 -0.017 0.98665

---

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

Residual standard error: 0.1129 on 388 degrees of freedom

Multiple R-squared: 0.7064, Adjusted R-squared: 0.7041

F-statistic: 311.1 on 3 and 388 DF, p-value: < 2.2e-16

Заметим незначительное улучшение исправленного коэффициента детерминации (0.7041), что говорит об улучшении модели.

|  |  |  |
| --- | --- | --- |
| **Коэффициенты уравнения множественной линейной регрессии, вычисленные** | | |
|  | **по явным формулам** | **c помощью функции lm** |
|  | 0.6685814608 | 0.668581 |
|  | -0.2324253516 | -0.232425 |
|  | -0.5430644400 | -0.543064 |
|  | -0.0009229535 | -0.000923 |

Подадим на вход модели 10 значений из обучающей выборки:

1) c(0.18478261, 0.14544939, 0.53571429)

2) c(0.64673913, 0.75446555, 0.23809524)

3) c(0.41304348, 0.33853133, 0.32738095)

4) c(0.01086957, 0.10547207, 0.80357143)

5) c(0.13586957, 0.17295152, 0.50595238)

6) c(0.72826087, 0.58151403, 0.17857143)

7) c(0.13043478, 0.13070598, 0.68452381)

8) c(0.21739130, 0.17210094, 0.35714286)

9) c(0.07608696, 0.05301956, 0.50000000)

10) c(0.03804348, 0.05160193, 0.55952381)

|  |  |  |
| --- | --- | --- |
| № | Истинное значение переменной отклика | Прогноз, полученный с помощью моделей |
| 1 | 0.42553191 | 0.5461505 |
| 2 | 0.10638298 | 0.1083197 |
| 3 | 0.29255319 | 0.3884332 |
| 4 | 0.90691489 | 0.6080353 |
| 5 | 0.42553191 | 0.5426111 |
| 6 | 0.05319149 | 0.1833508 |
| 7 | 0.55851064 | 0.5666516 |
| 8 | 0.37234043 | 0.5242627 |
| 9 | 0.72074468 | 0.6216424 |
| 10 | 0.63829787 | 0.6311996 |

**Выводы:**

Таким образом, коэффициенты уравнения множественной линейной регрессии, вычисленные двумя способами, совпадают.

Исправленный коэффициент детерминации равен 0.7041, что свидетельствует о достаточно точном прогнозе. Спрогнозированные значения подтверждают точность прогноза.

Отметим, что наибольшее влияние на расход топлива оказывает вес автомобиля. Факторы “мощность”, “вес” влияют на понижение расхода топлива, а фактор “ускорение” – на повышение. Также заметим, что фактор acceleration не влияет на результат прогноза.

**Код программы:**

setwd("C://Users/pc/Documents/Магистратура/2 семестр/R/task2 МЛР1")

getwd()

Непосредственное вычисление коэффициентов множественной линейной регрессии:

N=392 # Число наблюдений (автомобилей)

M=5 # Число факторов

# Создадим "шаблоны" матриц

Data = matrix(1 : (N\*(M+1)), ncol=(M+1)) # Данные из файла

X = matrix(1 : (N\*M), ncol=M) # Матрица значений факторов, расширенная единичным столбцом

Y = matrix(1 : N, ncol=1) # Столбец значений переменной отклика (mpg)

T = matrix(1 : (M+1)\*(M+1), ncol = (M+1) )

Teta = matrix(1 : (M+1), ncol = 1) # Вектор коэффициентов ЛРМ

# Прочтём файл с характеристиками автомобилей и расходом топлива и запишем данные в заготовленную матрицу

Data = read.table("data1.txt", sep = '\t')

# Нормализуем данные

for(i in 0:M+1)

{

myMax = max(Data[i])

myMin = min(Data[i])

Data[i] = (Data[i] - myMin)/(myMax - myMin)

}

#удалим первый и второй фактор

Data = Data[-2]

Data = Data[-1]

Data

M = 3

# Разделим значения факторов и переменной отклика

X = Data[, 1:M]

Y = Data[, (M+1)]

# Cоздадим вектор из единиц (как набор из N единиц)

odin=vector(length=N,mode='numeric')

odin = rep(1,N)

# Заготовим матрицу нужного размера

X1 = matrix(1 : (N\*(M+1)),ncol=(M+1))

# Заполним первый столбец матрицы X1 единицами

X1[,1] = odin

X1[,1]

# Остальные столбцы возьмём из матрицы Х

for(i in 1:M) X1[,i+1] = X[,i]

# Найдём вектор параметров множественной линейной регрессии Teta

# Вычислим сначала матрицу t(X1)

T1 = t(X1)

T1

# Вычислим матрицу T=t(X1)%\*%X1

T=T1%\*%X1

T

# Подключим библиотеку для работы с матрицами

library(MASS)

# Найдём обратную матрицу для Т

obr = ginv(T)

obr

# Вычислим вектор Teta - вектор коэффициентов линейной модели

Teta=obr%\*%T1%\*%Y

Teta=Teta[,1]

print(Teta)

# Введём данные нашего автомобиля (чтобы оценить его расход топлива)

myAuto = c(1, 0.41304348, 0.33853133, 0.32738095)

myAuto

myX=matrix(myAuto,ncol=M+1,byrow=TRUE)

tmyX=t(myX)

# Спрогнозируем расход топлива нашего автомобиля по нашей модели

myMpg=Teta%\*%tmyX

print(myMpg)

Построение уравнения регрессии с помощью функции lm:

# Прочтём файл с характеристиками автомобилей и расходом топлива

myData = read.csv("data1.txt", sep = '\t', header = TRUE)

myData

M = 5 # число факторов

# Нормализуем данные

for(i in 0:M+1)

{

myMax = max(myData[i])

myMin = min(myData[i])

myData[i] = (myData[i] - myMin)/(myMax - myMin)

}

myData

# вычислим попарные коэффициенты корреляции между предикторами

rm = cor(myData)

rm

# удалим первый и второй фактор

myData = myData[-2]

myData = myData[-1]

myData

# вычислим попарные коэффициенты корреляции между предикторами

rm = cor(myData)

rm

# Построим регрессионную модель

mymodel = lm(mpg ~ horsepower + weight + acceleration, data = myData)

mymodel

summary(mymodel)

# Введём данные нашего автомобиля (чтобы оценить его расход топлива)

myAuto=c(0.41304348, 0.33853133, 0.32738095)

myAuto=data.frame(t(myAuto))

colnames(myAuto)<-

c("horsepower", "weight", "acceleration")

myAuto

# Спрогнозируем расход топлива нашего автомобиля по нашей модели

newMpg = predict(mymodel, myAuto)

newMpg