* **Название и номер задания:**

Задание МЛР1 – нахождение параметров множественной линейной регрессии. Прогнозирование.

* **Постановка задачи:**

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

* Подберите данные для задачи.
* Постройте с помощью пакета R уравнение линейной регрессии (сделайте это 2мя способами: по явным формулам; с помощью функции lm).
* Сравните результаты с помощью таблицы.
* Приведите в отчёте матрицу попарных корреляций и сводную информацию (summary) модели.
* Проанализируйте степень влияния факторов на переменную отклика согласно каждой из полученных моделей.
* 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.
* Проанализируйте и объясните полученные результаты.
* Опишите полученные результаты в файле формата Word и прикрепите файл здесь.
* **Студент:**

Санамян Артак Размикович

* **Номер группы:**

09-715(1)

* **Ссылка на источник данных:**

<https://drive.google.com/open?id=1uMY5eq2yUjuV6WnH_pZrgXtczM9ueYLN>

* **Постановка задачи**

Общее назначение множественной регрессии состоит в анализе связи между несколькими независимыми переменными (называемыми также регрессорами или предикторами) и зависимой переменной.

Необходимо проанализировать зависимость между некоторыми факторами (рейтинг ресторана, расстояние до центра и процент свободных столиков) и итоговой стоимостью обеда.

* **Описание выполнения работы**

Необходимо найти коэффициенты уравнения множественной линейной регрессии, вычисленные двумя способами: по явным формулам и при помощи функции lm.

В таблице ниже приведены найденные коэффициенты.

|  |  |  |
| --- | --- | --- |
| **Коэффициенты уравнения множественной линейной регрессии, вычисленные** | | |
|  | **По явным формулам** | **С помощью функции lm** |
| Beta\_0 | 1050.89105 | 1050.891 |
| Beta\_1 | 343.07839 | 343.078 |
| Beta\_2 | -58.93371 | -58.934 |
| Beta\_3 | -108.64757 | -108.648 |

Регрессионные коэффициенты, найденные различными способами, совпадают с точностью до десятитысячных.

Приведем матрицу попарных корреляций и сводную информацию модели.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Rating** | **Distance** | **Free** | **Price** |
| **Rating** | 1.000000000 | 0.006547014 | 0.01477140 | 0.7274301 |
| **Distance** | 0.006547014 | 1.000000000 | 0.03195295 | -0.4055210 |
| **Free** | 0.014771399 | 0.031952955 | 1.00000000 | -0.3066665 |
| **Price** | 0.727430142 | -0.405521045 | -0.30666645 | 1.0000000 |

Рассматривая коэффициенты попарных корреляций, можно заметить, что два из трех факторов влияют на переменную отклика (цену) обратно пропорционально, то есть увеличение такого фактора приводит к снижению итогового результата. Иными словами, чем больше расстояние от ресторана до центра, и чем больше процент свободных столов, тем меньше итоговая стоимость обеда.

Помимо обратно пропорциональных факторов, в модели присутствует прямо пропорциональный фактор, который оказывает максимальное влияние на переменную отклика (рейтинг ресторана).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Сводная информация модели** | | | | | |
| **Информация о распределении остатков** | | | | | |
| **Минимум** | **Первая квартиль** | **Медиана** | | **Третья квартиль** | **Максимум** |
| -405.26 | -118.84 | -20.07 | | 92.44 | 991.97 |
| **Информация о коэффициентах** | | | | | |
|  | **Метод наименьших квадратов** | **Стандартные** **ошибки** | | **Значения t-статистики** | **p-значения** |
| **Price** | 1050.891 | 20.941 | | 50.18 | <2e-16 |
| **Rating** | 343.078 | 4.789 | | 71.64 | <2e-16 |
| **Distance** | -58.934 | 1.509 | | -39.06 | <2e-16 |
| **Guests** | -108.648 | 3.657 | | -29.71 | <2e-16 |
| **Значение статистики R-квадрат** | | | **Значение скорректированного R-квадрат** | | |
| 0.7902 | | | 0.7899 | | |

Сделаем прогноз итоговой стоимости обеда: на вход модели подадим двадцать значений из обучающей выборки и сравним результат с истинной стоимостью.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Rating** | **Distance** | **Guests** | **Price** | **Price\_Model** | **Difference** |
| 4.116562 | 3.0135618 | 2 | 2082.6128 | 2068.2991 | 14.31377 |
| 2.291105 | 7.1728650 | 1 | 1323.3322 | 1305.5484 | 17.78386 |
| 3.863138 | 8.2831317 | 3 | 1543.5746 | 1562.1518 | 18.57720 |
| 2.291893 | 6.3979771 | 2 | 1263.5586 | 1242.8382 | 20.72035 |
| 2.461222 | 4.0701044 | 1 | 1573.7810 | 1546.7690 | 27.01194 |
| 4.424685 | 6.5832806 | 2 | 1994.7593 | 1963.6327 | 31.12663 |
| 3.832502 | 6.7907047 | 3 | 1672.9204 | 1639.5955 | 33.32492 |
| 2.717108 | 4.7852923 | 3 | 1341.3922 | 1375.1143 | 33.72202 |
| 2.289318 | 4.8331737 | 4 | 1081.9691 | 1116.8793 | 34.91017 |
| 2.233595 | 8.2283173 | 4 | 860.6475 | 897.6738 | 37.02625 |
| 2.942564 | 8.9099574 | 3 | 1252.4304 | 1209.3815 | 43.04887 |
| 4.798278 | 6.2387727 | 1 | 2264.1517 | 2220.7550 | 43.39674 |
| 3.261017 | 6.3477432 | 3 | 1513.5758 | 1469.6368 | 43.93896 |
| 2.939891 | 3.9492602 | 4 | 1345.5080 | 1392.1693 | 46.66129 |
| 4.927391 | 7.3127758 | 4 | 1824.0458 | 1875.8132 | 51.76741 |
| 4.545981 | 6.1139596 | 4 | 1763.0018 | 1815.6102 | 52.60842 |
| 4.508773 | 6.2473165 | 4 | 1724.8303 | 1794.9858 | 70.15558 |
| 4.342473 | 3.8775882 | 3 | 1915.1563 | 1986.2362 | 71.07990 |
| 2.400525 | 1.3671050 | 3 | 1393.5701 | 1467.9479 | 74.37782 |
| 3.135378 | 5.2879483 | 2 | 1521.6373 | 1597.6379 | 76.00056 |

Исходя из полученных результатов, можно сказать, что прогнозируемая стоимость отличается от истинной стоимости на удовлетворительную величину (минимальная истинная стоимость обеда в данной таблице 861 у.е., а максимальная разница между истинной и прогнозируемой стоимостью 76 у.е.).

* **Выводы**

Мы рассмотрели влияние трех факторов (рейтинг ресторана, расстояние до центра и процент свободных столов) на стоимость обеда в целом. Получены две модели (с использованием стандартной функции lm() и по явным формулам), коэффициенты которых совпали. Исходя из таблицы попарных корреляций, стоит заметить, что не все факторы одинаково влияют на формирование стоимости: рейтинг ресторана является определяющим фактором, а расстояние до центра и процент свободных столов – обратно пропорциональные факторы.

Так же, сравнив на последнем этапе выполнения работы истинные и прогнозируемые значения переменной отклика, мы можем сказать, что полученные коэффициенты модели множественной линейной регрессии удовлетворяют ожидания, так как максимальное отклонение цены мало по сравнению со значением переменой отклика.

* **Приложение**

rm (list=ls())

#количество наблюдений

N = 2000

#количество признаков

M = 3

#создаем заготовки для будущих матриц

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

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

Y = matrix(1 : N, ncol=1)

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

Teta = matrix(1 : (M+1), ncol = 1)

#считываем данные

A=scan("aaa2.txt")

Data=matrix(A,ncol=M+1,byrow=TRUE)

#разделяем признаки и результат

X = Data[, 1:M]

Y = Data[, (M+1)]

#создадим вектор из единиц

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

odin = rep(1,N)

#задаем матрицу, дополненную единичным столбцом

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

X1[,1] = odin

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

T1 = t(X1)

T=T1%\*%X1

#подключим доп библиотеку

library(MASS)

#найдем обратную матрицу

obr = ginv(T)

#найдем коэффициенты первым способом

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

Teta=Teta[,1]

#преобразуем таблицу в матрицу

B = as.data.frame.matrix(Data)

#найдем коэффициенты вторым способом

mymodel = lm(V4 ~ V1 + V2 + V3, data = B)

summary(mymodel)

#распечатаем регрессионные коэффициенты

print(Teta)

#найдем коэффициенты корреляции

cor(B[,1],B[,4])

cor(B[,2],B[,4])

cor(B[,3],B[,4])

#получим сравнение фактической и ожидаемой стоимости

S = 6

W = 20

Rez = matrix(1 : (W\*S), ncol=S)

for(i in 1:20){

mycost=c(B[i,1],B[i,2],B[i,3])

mycost=data.frame(t(mycost))

colnames(mycost)<- c("V1","V2","V3")

Rez[i] <- B[i,1]

Rez[20+i] <- B[i,2]

Rez[40+i] <- B[i,3]

Rez[60+i] <- B[i,4]

Rez[80+i] <- predict(mymodel, mycost)

Rez[100+i] <- Mod(Rez[60+i]-Rez[80+i])

}