**Модульное задание 2**

Срок сдачи: 12 декабря 23:59. Максимум – 10 баллов.

**Задание 1**

По данным по Великобритании о потреблении цыплят (Y), среднедушевом доходе (X1), стоимости одного фунта цыплят (X2), стоимости одного фунта свинины (X3) и стоимости одного фунта говядины (X4) а) Необходимо построить, сравнить и проинтерпретировать уравнения регрессии вида:

1) (функция спроса)

2) (функция потребления)

3) (функция спроса-потребления)

4) (функция спроса с учетом цены на товары-заменители)

b) Применить тест Шапиро-Уилка для проверки нормальности распределения X1, X2, X3, X4.

с) Для данных из пункта b, которые не прошли проверку, применить преобразования Бокса-Кокса.

d) Применить критерий Манна-Уитни для оценки различий данных X3 и X4. Дать интерпретацию.

e) Применить критерий Флигнера-Клипера для X2 и X3. Применить критерий Стьюдента для этих данных с учетом полученных результатов. Привести интерпретацию.

Исходные данный для исследования приведены в табл. 1

Таблица 1

Исходные данные

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **t** | **Y** | **X1** | **X2** | **X3** | **X4** |
| 1 | 31,2 | 492,9 | 37,3 | 54,7 | 77,4 |
| 2 | 33,3 | 528,6 | 38,1 | 63,7 | 80,2 |
| 3 | 35,6 | 560,3 | 39,3 | 69,8 | 80,4 |
| 4 | 36,4 | 624,6 | 37,8 | 65,9 | 83,9 |
| 5 | 36,7 | 666,4 | 38,4 | 64,5 | 85,5 |
| 6 | 38,4 | 717,8 | 40,1 | 70 | 93,7 |
| 7 | 40,4 | 768,2 | 38,6 | 73,2 | 106,1 |
| 8 | 40,3 | 843,3 | 39,8 | 67,8 | 104,8 |
| 9 | 41,8 | 911,6 | 39,7 | 79,1 | 114 |
| 10 | 40,4 | 931,1 | 52,1 | 95,4 | 124,1 |
| 11 | 40,7 | 1021,5 | 48,9 | 94,2 | 127,6 |
| 12 | 40,1 | 1165,9 | 58,3 | 123,5 | 142,9 |
| 13 | 42,7 | 1349,6 | 57,9 | 129,9 | 143,6 |
| 14 | 44,1 | 1449,4 | 56,5 | 117,6 | 139,2 |
| 15 | 46,7 | 1575,5 | 63,7 | 130,9 | 165,5 |
| 16 | 50,6 | 1759,1 | 61,6 | 129,8 | 203,3 |
| 17 | 50,1 | 1994,2 | 58,9 | 128 | 219,6 |
| 18 | 51,7 | 2258,1 | 66,4 | 141 | 221,6 |
| 19 | 52,9 | 2478,7 | 70,4 | 168,2 | 232,6 |

**Задание 2**

По данным из файла «задание2.xlsx» показателей деятельности 50 фирм, оказывающих фото- и видео-услуги в городе N, за 2019 г. рассмотреть возможности оценивания нелинейной, в частности полиномиальной, регрессии прибыльности вложений в оборудование (profitability, %) на показатели среднего числа сотрудников на один проект (staff), среднемесячного числа проектов (projects), доли расходов на рекламу и продвижение своего бренда (adverts, %).

а) Найти степень полинома, при которой коэффициент детерминации оказывается близким к 1.

b) Показать, что результат достигнут за счет переобучения модели.

с) Выбрать оптимальную степень полинома, в том числе с содержательной точки зрения, предполагая, что фирмы можно считать однородными по прочим возможным показателям их деятельности.

**Задание 3**

Применить метод главных компонент для понижения размерности данных о результативности деятельности российских вузов. Данные в файле «задание3.xlsx» а) Определить минимальное количество компонент, которые необходимо использовать для сохранения 75% первоначальной информации. b) Выписать формулы зависимости главных компонент из пункта а) от первоначальных данных.

**Задание 4**

По данным из файла «задание4.xls» обучить модель логистической регрессии с целью прогнозирования увольнения сотрудника. а) Отобрать переменные, которые будут включаться в модель. Обосновать сделанный выбор. Обучить модель логистической регрессии. Дать интерпретацию полученных результатов. b) Привести матрицу ошибок, рассмотреть различные метрики качества модели. c) Построить ROC-кривую.

Описание переменных:

target – сотрудник уволится в следующем году (0-нет, 1-да)

age-возраст

educ- образование (1-начальное, 2-среднее, 3 -неоконченное высшее, 4-два и более высших образования, 5-кандидат наук, 6-доктор наук)

work\_exp – стаж работы (лет)

interest – оценка сотрудником интереса к работе (от 0 до 10)

coffee- среднее количество чашек кофе, выпиваемых сотрудником за рабочий день

boss\_educ -образование руководителя (1-начальное, 2-среднее, 3 -неоконченное высшее, 4-два и более высших образования, 5-кандидат наук, 6-доктор наук)

passport- наличие загранпаспорта у сотрудника (1-есть, 0 -нет)

green -степень озеленения офиса (от 1 до 5)

floor – этаж работы

children – количество детей у сотрудника

climate- удовлетворенность рабочей атмосферой (от 0 до 4)

offhour – наличие переработок (1-да, 0 -нет)

dist – расстояние до работы

salary – удовлетворенность заработной платой (1-да, 0 -нет)

height- рост сотрудника

lunch- средняя цена бизнес-ланча в офисе