Описание данных

Н.В. Артамонов

24 марта 2023 г.

Содержание

В этом разделе приведены описания основных наборов данных, используемых для решения задач.

Для работы с нужным датасетом необходимо загрузить его командой

```
data(dataset, package)
```

1 Кросс-секционные данные

Таблица 1: Набор данных `Labour` из пакета `Ecdat` (569 наблюдений) с данными о бельгийских фирмах за 1996 г.

capital	капитал (в млн евро)
labour	число сотрудников
output	выпуск (в млн евро)
wage	зарплата на одного сотрудника (в тыс евро)

Таблица 2: Набор данных `Electricity` из пакета `Ecdat` (158 наблюдений) с данными о производителях электроэнергии в US за 1970 г.

cost	общие издержки за год
q	общий выпуск электроэнергии
pl	уровень зарплата (wage rate)
pk	цена привлечения капитала (capital price index)
pf	цена на топливо (fuel price)

Таблица 3: Набор данных `sleep75` из пакета `wooldridge` (706 наблюдений). Основные переменные. Источник данных [5]

age	возраст (в годах)
educ	уровень образования (в года)
inlf	бинарная, 1 если участник рынка труда
leis1	нерабочее время, sleep-totwrk
smsa	бинарная, 1 если живёт в мегаполисе
male	гендерный фактор (бинарная, 1 если мужчина)
marr	семейный статус (бинарная, 1 если женат/замужем)
prot	бинарная, 1 если протестант
selfe	бинарная, 1 если самозанятый
sleep	продолжительность сна (мин/нед)
south	географический фактор (бинарная, 1 если живёт на юге)
spsepay	доход супруга/супруги
spwrk75	бинарная, 1 если супруг(а) работает
totwrk	занятость (мин/нед)
union	бинарная, 1 если член профсоюза
yngkid	бинарная, 1 если сеть дети младше 3 лет
yrsmarr	сколько лет женат/замужем
hrwage	почасовая оплата

Таблица 4: Набор данных `wage2` из пакета `wooldridge` (935 наблюдений). Основные переменные. Источник данных [6]

wage	месячная зарплата
hours	недельная занятость в часах
IQ	результаты теста IQ
KWW	результаты теста knowledge of world work
educ	уровень образования (в годах)
exper	опыт работы в годах
tenure	стаж работы на текущем месте
age	возраст (в годах)
married	семейный статус (бинарная, 1 если женат/замужем)
south	географический фактор (бинарная, 1 если живёт на юге)
urban	место жительства (1 если живет в городе)
sibs	число братьев/сестёр
brthord	какой по счёту ребёнок в семье
meduc	уровень образования матери (в годах)
feduc	уровень образования отца (в годах)

Таблица 5: Набор данных `wage1` из пакета `wooldridge` (526 наблюдений). Основные переменные.

wage	средняя почасовая оплата
educ	уровень образования (в годах)
female	гендерный фактор
exper	опят работы
tenure	стаж на текущем месте работы
married	семейный статус
smsa	живёт ли в мегаполисе (бинарная)
south	географический фактор (бинарная)
west	географический фактор (бинарная)
northcen	географический фактор (бинарная)

Таблица 6: Набор данных `loanapp` из пакета `wooldridge` (1989 наблюдений). Основные переменные. Источник данных [9]

approve	бинарная, 1 если кредитная заявка одобрена
appinc	доход заявителя (в \$1000)
mortno	бинарная, 1 если нет ипотечной кредитной истории
unem	уровень безработицы в отрасли в %
dep	количество иждивенцев
male	гендерный фактор
married	семейный статус
yjob	стаж на текущей работе
self	бинарная, 1 если самозанятый

Таблица 7: Набор данных `SwissLabor` о рынке труда Швейцарии из пакета `AER` (872 наблюдений). Источник данных [7]

participation	Является ли участником рынка труда? (фактор, ''yes''/''no'')
income	логарифм дополнительного дохода (nonlabor income)
age	возраст (в десятилетиях)
education	уровень образования
youngkids	число маленьких детей (младше 7 лет)
oldkids	число старшийх детей (страше 7 лет)
foreign	является ли иностранцем? (фактор, ''yes''/''no'')

Таблица 8: Набор данных из файла `Mroz` из пакета `Ecdat` содержит данные о рынке труда замужних женщин. Основные переменные. Источник данных [10]

LFP	бинарная, 1 женщина работает
WHRS	уровень занятости (в часах)
KL6	число детей моложе 6 лет в семье
K618	число детей от 6 до 18 лет в семье
WA	возраст
WE	уровень образования (в годах)
WW	средняя почасовая оплата
HHRS	занятость мужа
HA	возраст мужа
HE	уровень образования мужа (а годах)
HW	зарплата мужа
FAMINC	доход домашнего хозяйства
WMED	уровень образования матери
WFED	уровень образования отца
UN	уровень безработицы в стране проживания
CIT	бинарная, 1 если живет в мегаполисе
AX	предыдущий опыт работы (в годах)

Таблица 9: Набор данных `diamonds` из пакета `ggplot2` с данными о бриллиантах (53940 наблюдений). Основные переменные.

price	цена бриллианта
carat	вес бриллианта (в каратах)
cut	качество огранки (упорядоченный фактор с уровнями Fair<Good< Very Good<Premium<Ideal)
color	цвет (упорядоченный фактор с уровнями J<I<H<G<F<E<D
clarity	прозрачность (упорядоченный фактор с уровнями I1<SI2<SI1<VS2<VS1<VVS2<VVS1<IF
x	длина (в мм)
y	ширина (в мм)
z	глубина (в мм)

Таблица 10: Набор данных `Diamond` из пакета `Ecdat` с данными о бриллиантах (308 наблюдений). Основные переменные. Источник данных [4]

carat	вес бриллианта (в каратах)
colour	цвет (фактор с уровнями D,E,F,G,H,I)
clarity	прозрачность (фактор с уровнями IF,VVS1,VVS2,VS1,VS2)
certification	орган по сертификации (фактор с уровнями GIA,IGI,HRD)
price	цена в Сингапуре

Таблица 11: Набор данных `nlsw88` из пакета `Counterfactual` о занятости женщин в США (2246 наблюдений). Основные переменные. Источник данных: сайт Stata http://www.stata-press.com/data/r10/g.html

hours	недельная занятость (в часах)
married	семейный статус
ttl_exp	общий стаж работы
smsa	бинарная, 1 если живёт в мегаполисе
south	географический фактор (бинарная, 1 если живёт на юге)
wage	почасовая оплата (в $)
age	возраст (в годах)
grade	уровень образования (в годах)

Таблица 12: Набор данных из файла `default.csv` с сайта http://meit.mgimo.ru/node/237 о банкротствах по студенческим займам (6778 наблюдений)

default	бинарная переменная, равная 1 если индивид признал себя банкротом по студенческому займу
age	возраст
adepcnt	количество иждивенцев у индивида плюс 1
acadmos	количество месяцев, которые индивид прожил по текущему адресу
majordrg	количество зарегистрированных серьёзных правонарушений у этого индивида
minordrg	количество зарегистрированных мелких правонарушений у этого индивида
ownrent	1, если индивид живёт в собственном дома, и 0, если снимает
income	месячный доход в $
spending	среднемесячный расход по кредитной карте
inc_rep	income, делённая на количество иждивенцев
exp_inc	доля месячных расходов по кредитной карте в годовой заработной плате
selfempl	1, если индивид самозанятый, и 0 иначе

Таблица 13: Набор данных `stockton3` из пакета `PoEdata` с данными о стоимости домов (2610 наблюдений). Основные переменные. Источник данных https://github.com/ccolonescu/PoEdata

sprice	цена продажи дома (в $)
livarea	жилая площадь (кв.футы)
pool	наличие бассейна (бинарная)
lgelot	размер участка (бинарный фактор, 1 если участок больше 5 акров)
age	возраст (в годах)
beds	число спален

Таблица 14: Набор данных из файла `applications.csv` о поступивших на магистерские и PhD-программы (400 наблюдений)

admit	бинарный фактор, 1 если заявка одобрена
GPA	средняя оценка за время обучения
GRE	балл за экзамен graduate record exam
rank	категориальная переменная, обозначающая престиж университета (1 – высший престиж, 4 – низший престиж)

Таблица 15: Набор данных `Consumption` из пакета `Ecdat` об индивидуальных доходах и расходах в Канаде (квартальные данные с 1947Q1 по 1996Q4).

yd	индивидуальный располагаемый доход в ценах 1986
ce	индивидуальные расходы на потребление в ценах 1986

2 Временные ряды

Таблица 16: Набор данных `Icecream` из пакета `Ecdat` о потреблении мороженого в США (недельные данные с 1951–03–18 по 1953–07–11, всего 30 наблюдений). Источник данных [8]

cons	потребление мороженого (в пинтах)
income	средний недельный доход семьи (в $)
price	цена мороженого (за пинту)
temp	средняя температура (по Фаренгейту)

Таблица 17: Панель `Guns` из пакета `AER` с данными по 51 штату США с 1977 по 1999 гг. (всего 1173 наблюдения). Основные переменные. Источник данных [1]

state	factor indicating state
year	factor indicating year
violent	violent crime rate (incidents per 100,000 members of the population)
murder	murder rate (incidents per 100,000).
robbery	robbery rate (incidents per 100,000)
prisoners	incarceration rate in the state in the previous year (sentenced prisoners per 100,000 residents; value for the previous year)
afam	percent of state population that is African-American, ages 10 to 64
cauc	percent of state population that is Caucasian, ages 10 to 64
male	percent of state population that is male, ages 10 to 29
population	state population, in millions of people.
income	real per capita personal income in the state (US dollars)
density	population per square mile of land area, divided by 1,000
law	factor. Does the state have a shall carry law in effect in that year?

3 Панельные данные

Таблица 18: Панель `LaborSupply` из пакета `plm, Ecdat` с данными по 532 индивидуумам с 1979 по 1988 гг. (всего 5320 наблюдений). Основные переменные. Источник данных [11]

lnhr	логарифм годовой занятости в часах
lnwg	логарифм почасовой оплаты
kids	число детей
age	возраст
disab	бинарная, 1 если плохое здоровье

Таблица 19: Панель `Cigar` из пакета `plm` с данными по 46 штатам США с 1963 по 1992 гг. (всего 1380 наблюдений). Основные переменные. Источник данных [2]

state	state abbreviation
year	the year
price	price per pack of cigarettes
pop	population
pop16	population above the age of 16
cpi	consumer price index (1983=100)
ndi	per capita disposable income
sales	cigarette sales in packs per capita
pimin	minimum price in adjoining states per pack of cigarettes

Таблица 20: Панель `Gasoline` из пакета `plm, Ecdat` с данными о потреблении бензина по 18 странам OECD с 1960 по 1978 гг. (всего 342 наблюдений). Основные переменные. Источник данных [3]

lgaspcar	логарифм потребления бензина
lincomep	логарифм реального дохода на душу населения
lrpmg	логарифм реальной цены на бензин
lcarpcap	логарифм объёма рынка машин

Таблица 21: Панель `Loan` Dream Housing Finance company deals in all home loans. They have a presence across all urban, semi-urban, and rural areas. Customer-first applies for a home loan after that company validates the customer eligibility for a loan.(всего 342 наблюдений). Основные переменные. Источник данных Loan

LoanID	Unique Loan ID
Gender	Male/ Female
Married	applicant married (Y/N)
Dependents	Number of dependents
Education	Applicant Education (Graduate/ Under Graduate)
Self Employed	Self-employed (Y/N)
ApplicantIncome	ApplicantIncome
CoapplicantIncome	CoapplicantIncome
LoanAmount	LoanAmount
Loan Amount Term	Loan Amount Term
Credit History	credit history meets guidelines
Property Area	Urban/ Semi-Urban/ Rural
Loan Status	Loan approved

Таблица 22: `1Custumers.csv` Shop Customer Data is a detailed analysis of a imaginative shop's ideal customers. It helps a business to better understand its customers. The owner of a shop gets information about Customers through membership cards. Dataset consists of 2000 records and 8 columns.

Customer ID	—
Gender	—
Age	—
Annual Income	—
Spending Score	Score assigned by the shop, based on customer behaviour and spending nature
Profession	—
Work Experience	in years
Family Size	—

Таблица 23: `2laptops.csv` This is a dataset about laptops scrap from Flipkart which contains information on various aspects of laptops such as their price, discount, specifications, and warranty. The dataset contains a total of 920 entries, each representing a single laptop. It is important to note that there are some missing values in the dataset for some of the columns, including "discount "In build sw and "warranty". This should be taken into consideration when analyzing the data. Additionally, all the columns have a data type of "object which may require further processing to convert the data into a usable format.

title	provides a brief description of the laptop
price	includes the cost of the laptop
discount	mentions any applicable discounts on the laptop's price
Processor	specifies the type of processor used in the laptop
RAM	mentions the amount of RAM the laptop has
OS	lists the operating system installed on the laptop
SSD	indicates the size of the solid-state drive
Display	mentions the screen size and display specifications of the laptop
In build sw	lists any software pre-installed on the laptop
warranty	provides information on the warranty offered for the laptop

Таблица 24: `3insurance.csv` Machine Learning with R by Brett Lantz is a book that provides an introduction to machine learning using R. As far as I can tell, Packt Publishing does not make its datasets available online unless you buy the book and create a user account which can be a problem if you are checking the book out from the library or borrowing the book from a friend. All of these datasets are in the public domain but simply needed some cleaning up and recoding to match the format in the book.

age	age of primary beneficiary
sex	insurance contractor gender, female, male
bmi	Body mass index, providing an understanding of body, weights that are relatively high or low relative to height objective index of body weight (kg/m^2) using the ratio of height to weight, ideally 18.5 to 24.9
children	Number of children covered by health insurance Number of dependents
smoker	Smoking
region	the beneficiary's residential area in the US, northeast, southeast southwest, northwest.
charges	Individual medical costs billed by health insurance

Таблица 25: `4winequality-red.csv` The two datasets are related to red and white variants of the Portuguese "Vinho Verde"wine. For more details, consult the reference [Cortez et al., 2009]. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.). These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are much more normal wines than excellent or poor ones). This dataset is also available from the UCI machine learning repository,winequality Content

fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol
Output variable (based on sensory data)
quality (score between 0 and 10)

Таблица 26: `5marketing-campaign.csv` Customer Personality Analysis is a detailed analysis of a company's ideal customers. It helps a business to better understand its customers and makes it easier for them to modify products according to the specific needs, behaviors and concerns of different types of customers. Customer personality analysis helps a business to modify its product based on its target customers from different types of customer segments. For example, instead of spending money to market a new product to every customer in the company's database, a company can analyze which customer segment is most likely to buy the product and then market the product only on that particular segment. Content:

People

ID	Customer's unique identifier
YearBirth	Customer's birth year
Education	Customer's education level
Marital Status	Customer's marital status
Income	Customer's yearly household income
Kidhome	Number of children in customer's household
Teenhome	Number of teenagers in customer's household
Dt Customer	Date of customer's enrollment with the company
Recency	Number of days since customer's last purchase
Complain	1 if the customer complained in the last 2 years, 0 otherwise

Products

MntWines	Amount spent on wine in last 2 years
MntFruits	Amount spent on fruits in last 2 years
MntMeatProducts	Amount spent on meat in last 2 years
MntFishProducts	Amount spent on fish in last 2 years
MntSweetProducts	Amount spent on sweets in last 2 years
MntGoldProds	Amount spent on gold in last 2 years

Promotion

NumDealsPurchases	Number of purchases made with a discount
AcceptedCmp1	1 if customer accepted the offer in the 1st campaign, 0 otherwise
AcceptedCmp2	1 if customer accepted the offer in the 2nd campaign, 0 otherwise
AcceptedCmp3	1 if customer accepted the offer in the 3rd campaign, 0 otherwise
AcceptedCmp4	1 if customer accepted the offer in the 4th campaign, 0 otherwise
AcceptedCmp5	1 if customer accepted the offer in the 5th campaign, 0 otherwise
Response	1 if customer accepted the offer in the last campaign, 0 otherwise

Place

NumWebPurchases	Number of purchases made through the company's website
NumCatalogPurchases	Number of purchases made using a catalogue
NumStorePurchases	Number of purchases made directly in stores
NumWebVisitsMonth	Number of visits to company's website in the last month

Таблица 27: Панель `6PlacementDataFullClass.csv` This data set consists of Placement data of students in a XYZ campus. It includes secondary and higher secondary school percentage and specialization. It also includes degree specialization, type and Work experience and salary offers to the placed students

Таблица 28: `californiaHousingPrices.csv` Context This is the dataset used in the second chapter of Aurélien Géron's recent book 'Hands-On Machine learning with Scikit-Learn and TensorFlow'. It serves as an excellent introduction to implementing machine learning algorithms because it requires rudimentary data cleaning, has an easily understandable list of variables and sits at an optimal size between being to toyish and too cumbersome. The data contains information from the 1990 California census. So although it may not help you with predicting current housing prices like the Zillow Zestimate dataset, it does provide an accessible introductory dataset for teaching people about the basics of machine learning. california Content

1. longitude	A measure of how far west a house is; a higher value is farther west
2. latitude	A measure of how far north a house is; a higher value is farther north
3. housingMedianAge	Median age of a house within a block; a lower number is a newer buildin
4. totalRooms	Total number of rooms within a block
5. totalBedrooms	Total number of bedrooms within a block
6. population	Total number of people residing within a block
7. households	Total number of households, a group of people residing within a home u
8. medianIncome	Median income for households within a block of houses (measured in ten
9. medianHouseValue	Median house value for households within a block (measured in US Dolla
10. oceanProximity	Location of the house w.r.t ocean/sea

Список литературы

[1] Ayres, I., and Donohue, J.J. (2003). Shooting Down the 'More Guns Less Crime' Hypothesis. Stanford Law Review, 55, 1193–1312

[2] Baltagi B, Levin D (1992). Cigarette taxation: Raising revenues and reducing consumption. Structural Change and Economic Dynamics, 3(2), 321-335

[3] Baltagi, B.H. and Y.J. Griggin (1983) "Gasoline demand in the OECD: an application of pooling and testing procedures", European Economic Review, 22.

[4] Chu, Singfat (2001) "Pricing the C's of Diamond Stones", Journal of Statistics Education, 9(2).

[5] J.E. Biddle and D.S. Hamermesh (1990), "Sleep and the Allocation of Time," Journal of Political Economy 98, 922-943.

[6] M. Blackburn and D. Neumark (1992), "Unobserved Ability, Efficiency Wages, and Interindustry Wage Differentials", Quarterly Journal of Economics 107, 1421-1436.

[7] Gerfin, M. (1996). Parametric and Semi-Parametric Estimation of the Binary Response Model of Labour Market Participation. Journal of Applied Econometrics, 11, 321–339.

[8] Hildreth, C. and J. Lu (1960) Demand relations with autocorrelated disturbances, Technical Bul- letin No 2765, Michigan State University.

[9] W.C. Hunter and M.B. Walker (1996), "The Cultural Affinity Hypothesis and Mortgage Lending Decisions", Journal of Real Estate Finance and Economics 13, 57-70

[10] Mroz, T. (1987) "The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions", Econometrica, 55, 765-799.

[11] Ziliak, Jim (1997) "Efficient Estimation With Panel Data when Instruments are Predetermined: An Empirical Comparison of Moment-Condition Estimators", Journal of Business and Economic Statistics, 419-431