Workshop03.R

kensuzuki

2024-05-27

```
### Simple Linear Regression ###

###### Step 1 - Load the data ######

# Set the Working Directory
setwd("~/Desktop/R_Workshop/R Workshop Day3")
# This line of code changes based on your local file location.


# Load the data
dat <- read.csv("~/Desktop/R_Workshop/R Workshop Day3/Data.csv")


# Load the data from a pop-up window
# dat <- read.csv(file.choose())



####### Step 2 - Explore the data ######

# Check the first few rows
head(dat)
```

```
##     AGE ECON_SOCIAL_CUL ED_RESOURCE ENV_AWARENESS HOUSE_POS PARENT_ED REPEAT
## 1 15.92         -0.2161     -0.7218       -0.4698   -0.3972        12      0
## 2 15.50          0.1575     -0.7218        0.0940   -2.0061        15      0
## 3 15.75         -0.7320     -0.7218        0.5056   -0.2828        12      0
## 4 15.67          0.8365      1.1563        2.2983   -0.2711        15      0
## 5 15.50         -0.4236     -0.7218        1.0611   -1.8613        12      0
## 6 15.75         -1.0457      0.8021       -0.0230   -1.7548        12      1
##   SCI_KNOWLEDGE SELF_EFFICACY SEX LANGUAGE STU_SCI_LITERACY TECH_RESOURCE
## 1       -0.8645        0.1133   0        0          419.516       -0.0295
## 2       -0.5216       -0.2306   0        0          479.997       -2.0167
## 3        0.2847        0.4572   0        0          482.213        0.0203
## 4        0.8420        0.3155   0        0          402.681       -0.1135
## 5        0.1889        0.5560   0        0          574.040       -1.7086
## 6       -0.1933       -0.5935   0        0          509.639       -2.0427
##    WEALTH
## 1 -0.0025
## 2 -2.1628
## 3  0.5426
## 4 -0.2688
```

```
## 5 -2.0633
## 6 -2.8299
```

```
# Get the structure of the data
str(dat)
```

```
## 'data.frame':    99 obs. of  14 variables:
##  $ AGE              : num  15.9 15.5 15.8 15.7 15.5 ...
##  $ ECON_SOCIAL_CUL : num  -0.216 0.158 -0.732 0.837 -0.424 ...
##  $ ED_RESOURCE      : num  -0.722 -0.722 -0.722 1.156 -0.722 ...
##  $ ENV_AWARENESS    : num  -0.47 0.094 0.506 2.298 1.061 ...
##  $ HOUSE_POS        : num  -0.397 -2.006 -0.283 -0.271 -1.861 ...
##  $ PARENT_ED        : int  12 15 12 15 12 12 12 15 12 15 ...
##  $ REPEAT           : int  0 0 0 0 0 1 1 0 0 0 ...
##  $ SCI_KNOWLEDGE    : num  -0.865 -0.522 0.285 0.842 0.189 ...
##  $ SELF_EFFICACY    : num  0.113 -0.231 0.457 0.316 0.556 ...
##  $ SEX              : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ LANGUAGE         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ STU_SCI_LITERACY: num  420 480 482 403 574 ...
##  $ TECH_RESOURCE    : num  -0.0295 -2.0167 0.0203 -0.1135 -1.7086 ...
##  $ WEALTH           : num  -0.0025 -2.1628 0.5426 -0.2688 -2.0633 ...
```

```
# Summary statistics of the data
summary(dat)
```

```
##        AGE         ECON_SOCIAL_CUL     ED_RESOURCE       ENV_AWARENESS
##  Min.   :15.42   Min.   :-3.6745   Min.   :-4.38610   Min.   :-3.37650
##  1st Qu.:15.67   1st Qu.:-1.9112   1st Qu.:-1.44320   1st Qu.:-0.65315
##  Median :15.92   Median :-1.0983   Median :-0.72180   Median :-0.11130
##  Mean   :15.88   Mean   :-1.1622   Mean   :-0.82425   Mean   :-0.08519
##  3rd Qu.:16.17   3rd Qu.:-0.2645   3rd Qu.:-0.03935   3rd Qu.: 0.52825
##  Max.   :16.33   Max.   : 0.8365   Max.   : 1.15630   Max.   : 2.29830
##    HOUSE_POS         PARENT_ED         REPEAT       SCI_KNOWLEDGE
##  Min.   :-4.5158   Min.   : 3.00   Min.   :0.0000   Min.   :-2.7904
##  1st Qu.:-2.5907   1st Qu.: 9.00   1st Qu.:0.0000   1st Qu.:-0.7249
##  Median :-1.9627   Median :12.00   Median :1.0000   Median :-0.1933
##  Mean   :-1.9475   Mean   :11.27   Mean   :0.6465   Mean   :-0.1937
##  3rd Qu.:-1.3304   3rd Qu.:15.00   3rd Qu.:1.0000   3rd Qu.: 0.2492
##  Max.   : 0.2757   Max.   :15.00   Max.   :1.0000   Max.   : 2.1552
##  SELF_EFFICACY          SEX            LANGUAGE       STU_SCI_LITERACY
##  Min.   :-2.3710   Min.   :0.0000   Min.   :0.00000   Min.   :239.9
##  1st Qu.:-0.9128   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:353.6
##  Median :-0.3937   Median :1.0000   Median :0.00000   Median :394.6
##  Mean   :-0.2623   Mean   :0.5758   Mean   :0.09091   Mean   :408.0
##  3rd Qu.: 0.2752   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:458.5
##  Max.   : 3.2775   Max.   :1.0000   Max.   :1.00000   Max.   :600.0
##  TECH_RESOURCE         WEALTH
##  Min.   :-3.3808   Min.   :-4.9389
##  1st Qu.:-2.5873   1st Qu.:-2.5840
##  Median :-1.9553   Median :-2.0503
##  Mean   :-1.9074   Mean   :-1.9604
##  3rd Qu.:-1.2896   3rd Qu.:-1.4027
##  Max.   : 0.7936   Max.   : 0.8853
```

```r
####### Step 3 - Data Cleaning: MISSING DATA ########

## Important Note ##

# Missing values can be problematic in data analysis.
# Mishandling of missing data can lead to biased results and incorrect conclusions.

# There are several ways to handle missing values, such as deletion or imputation.
# For instance, if the missing values are random, deletion might be appropriate.
# If the missing values are systematic, such that they are related to other variables,
# imputation, which fills in the missing values with estimated values based on the observed data.

# Dealing with missing data is extensive and can easily take up an entire workshop on its own.
# For now, we will focus on simple deletion of missing values.
```

```r
# Check for missing values
any(is.na(dat))
```

```
## [1] FALSE
```

```r
# If there are missing values, you can remove rows with NA
dat_clean <- na.omit(dat)
```