hw1_code

Jiaqi Chen

2/13/2022

```
library(RNHANES)
library(tidyverse)
library(summarytools)
library(leaps)
```

(a) Fit a linear model using least squares on the training data. Is there any potential disadvantage of this model?

Import and clean data

```
housing_test = read.csv("./housing_test.csv") %>%
  janitor::clean_names()

housing_training = read.csv("./housing_training.csv") %>%
  janitor::clean_names()
```

Summary statistics of the predictors and the response:

```
st_options(plain.ascii = FALSE,
           style = "rmarkdown",
           dfSummary.silent = TRUE,
           footnote = NA,
           subtitle.emphasis = FALSE)

dfSummary(housing_training)
```

```
## ### Data Frame Summary
## **housing_training**
## **Dimensions:** 1440 x 26
## **Duplicates:** 0
##
## -----------------------------------------------------------------------------------------------------
## No   Variable            Stats / Values                   Freqs (% of Valid)     Graph
## ---- ------------------- -------------------------------- ---------------------- -------------------
## 1    gr_liv_area\        Mean (sd) : 1477.6 (484.9)\      838 distinct values    \ \ \ \ :\
##      [integer]           min < med < max:\                                       \ \ \ \ : :\
##                          492 < 1432.5 < 4316\                                    \ \ \ \ : :\
##                          IQR (CV) : 628.2 (0.3)                                  \ \ : : :\
##                                                                                  \ \ : : : : .
##
## 2    first_flr_sf\       Mean (sd) : 1134.2 (367.9)\      744 distinct values    \ \ : :\
##      [integer]           min < med < max:\                                       \ \ : :\
##                          372 < 1055 < 3228\                                      \ \ : : :\
##                         IQR (CV) : 475.8 (0.3)                                   \ \ : : : :\
##                                                                                  . : : : : .
##
## 3    second_flr_sf\     Mean (sd) : 338.6 (422.8)\        402 distinct values    :\
##      [integer]          min < med < max:\                                        :\
##                         0 < 0 < 1872\                                            :\
##                         IQR (CV) : 704.8 (1.2)                                   :\
##                                                                                  : \ \ . : : . .
##
## 4    total_bsmt_sf\     Mean (sd) : 1035.1 (413.8)\       695 distinct values    \ \ \ \ :\
##      [integer]          min < med < max:\                                        \ \ \ \ : .\
##                         0 < 975 < 3206\                                          \ \ \ \ : :\
##                         IQR (CV) : 469.8 (0.4)                                   \ \ \ \ : : :\
##                                                                                  . : : : : :
##
## 5    low_qual_fin_sf\   Mean (sd) : 4.7 (44.3)\           20 distinct values     :\
##      [integer]          min < med < max:\                                        :\
##                         0 < 0 < 697\                                             :\
##                         IQR (CV) : 0 (9.5)                                       :\
##                                                                                  :
##
## 6    wood_deck_sf\      Mean (sd) : 98.4 (133.8)\         290 distinct values    :\
##      [integer]          min < med < max:\                                        :\
##                         0 < 0 < 1424\                                            :\
##                         IQR (CV) : 169 (1.4)                                     : .\
##                                                                                  : : .
##
## 7    open_porch_sf\     Mean (sd) : 45.1 (63.5)\          187 distinct values    :\
##      [integer]          min < med < max:\                                        :\
##                         0 < 26 < 570\                                            :\
##                         IQR (CV) : 68 (1.4)                                      :\
##                                                                                  : : .
##
## 8    bsmt_unf_sf\       Mean (sd) : 520.6 (415)\          778 distinct values    :\
##      [integer]          min < med < max:\                                        : :\
##                         0 < 431.5 < 2336\                                        : : : .\
##                         IQR (CV) : 574.5 (0.8)                                   : : : :\
##                                                                                  : : : : : . .
##
## 9    mas_vnr_area\      Mean (sd) : 95.4 (168)\           295 distinct values    :\
##      [integer]          min < med < max:\                                        :\
##                         0 < 0 < 1600\                                            :\
##                         IQR (CV) : 149.2 (1.8)                                   :\
##                                                                                  : : .
##
## 10   garage_cars\       Mean (sd) : 1.8 (0.7)\            0 :  61 ( 4.2%)\       \
##      [integer]          min < med < max:\                 1 : 393 (27.3%)\       IIIII \
##                         0 < 2 < 5\                        2 : 820 (56.9%)\       IIIIIIIIIII \
##                         IQR (CV) : 1 (0.4)                3 : 158 (11.0%)\       II \
##                                                           4 :   7 ( 0.5%)\       \
##                                                           5 :   1 ( 0.1%)
##
## 11   garage_area\       Mean (sd) : 471.9 (201.6)\        442 distinct values    \ \ \ \ \ \ :\
##      [integer]          min < med < max:\                                        \ \ \ \ \ \ :\
##                         0 < 480 < 1356\                                       \ \ \ \ : : .\
##                         IQR (CV) : 240 (0.4)                                  \ \ . : : :\
##                                                                               . : : : : : .
##
## 12   year_built\        Mean (sd) : 1970.2 (29.4)\        107 distinct values   \ \ \ \ \ \ \ \ \ \
##      [integer]          min < med < max:\                                     \ \ \ \ \ \ \ \ \ \
##                         1872 < 1972 < 2009\                                   \ \ \ \ \ \ \ \ \ \
##                         IQR (CV) : 44 (0)                                     \ \ \ \ \ \ . \ \ .
##                                                                               \ \ \ \ . : : : : :
##
## 13   tot_rms_abv_grd\   Mean (sd) : 6.4 (1.5)\            3 :  17 ( 1.2%)\       \
##      [integer]          min < med < max:\                 4 : 102 ( 7.1%)\       I \
##                         3 < 6 < 12\                       5 : 309 (21.5%)\       IIII \
##                         IQR (CV) : 2 (0.2)                6 : 413 (28.7%)\       IIIII \
##                                                           7 : 321 (22.3%)\       IIII \
##                                                           8 : 158 (11.0%)\       II \
##                                                           9 :  62 ( 4.3%)\       \
##                                                           10 :  39 ( 2.7%)\      \
##                                                           11 :  11 ( 0.8%)\      \
##                                                           12 :   8 ( 0.6%)
##
## 14   full_bath\         Mean (sd) : 1.5 (0.5)\            0 :   3 ( 0.2%)\       \
##      [integer]          min < med < max:\                 1 : 683 (47.4%)\       IIIIIIIII \
##                         0 < 2 < 4\                        2 : 728 (50.6%)\       IIIIIIIIII \
##                         IQR (CV) : 1 (0.4)                3 :  25 ( 1.7%)\       \
##                                                           4 :   1 ( 0.1%)
##
## 15   overall_qual\      1\. Above_Average\                382 (26.5%)\           IIIII \
##      [character]        2\. Average\                      425 (29.5%)\           IIIII \
##                         3\. Below_Average\                104 ( 7.2%)\           I \
##                         4\. Excellent\                    42 ( 2.9%)\            \
##                         5\. Fair\                         22 ( 1.5%)\            \
##                         6\. Good\                         302 (21.0%)\           IIII \
##                         7\. Very_Excellent\               13 ( 0.9%)\            \
##                         8\. Very_Good                     150 (10.4%)            II
##
## 16   kitchen_qual\      1\. Excellent\                    72 ( 5.0%)\            I \
##      [character]        2\. Fair\                         24 ( 1.7%)\            \
##                         3\. Good\                         559 (38.8%)\           IIIIIII \
##                         4\. Typical                       785 (54.5%)            IIIIIIIIII
##
## 17   fireplaces\        Mean (sd) : 0.6 (0.7)\            0 : 705 (49.0%)\       IIIIIIIII \
##      [integer]          min < med < max:\                 1 : 616 (42.8%)\       IIIIIIII \
##                         0 < 1 < 3\                        2 : 113 ( 7.8%)\       I \
##                         IQR (CV) : 1 (1.1)                3 :   6 ( 0.4%)
##
## 18   fireplace_qu\      1\. Excellent\                    17 ( 1.2%)\            \
##      [character]        2\. Fair\                         41 ( 2.8%)\            \
##                         3\. Good\                         344 (23.9%)\           IIII \
##                         4\. No_Fireplace\                 705 (49.0%)\           IIIIIIIII \
##                         5\. Poor\                         25 ( 1.7%)\            \
##                         6\. Typical                       308 (21.4%)            IIII
##
## 19   exter_qual\        1\. Excellent\                    39 ( 2.7%)\            \
##      [character]        2\. Fair\                       15 ( 1.0%)\            \
##                         3\. Good\                       469 (32.6%)\           IIIIII \
##                         4\. Typical                     917 (63.7%)            IIIIIIIIIIII
##
## 20   lot_frontage\      Mean (sd) : 55 (32.5)\           105 distinct values    \ \ \ \ \ \ :\
##      [integer]          min < med < max:\                                       \ \ \ \ : :\
##                         0 < 60 < 174\                                           : \ \ : :\
##                         IQR (CV) : 39 (0.6)                                     : \ \ : : :\
##                                                                                 : : : : : .
##
## 21   lot_area\          Mean (sd) : 10101 (8302.1)\      1063 distinct values   :\
##      [integer]          min < med < max:\                                       :\
##                         1470 < 9306.5 < 164660\                                 :\
##                         IQR (CV) : 4187 (0.8)                                   :\
##                                                                                 :
##
## 22   longitude\         Mean (sd) : -93.6 (0)\           1411 distinct values   \ \ \ \ \ \ \ \ : .
##      [numeric]          min < med < max:\                                       : \ \ \ \ . : : : :\
##                         -93.7 < -93.6 < -93.6\                                  : . \ \ : : : : :\
##                         IQR (CV) : 0 (0)                                        : : : : : : : :\
##                                                                                 : : : : : : : :
##
## 23   latitude\          Mean (sd) : 42 (0)\              1400 distinct values   \ \ \ \ \ \ \ \ . \
##      [numeric]          min < med < max:\                                       \ \ \ \ \ \ \ \ : :
##                         42 < 42 < 42.1\                                         \ \ \ \ \ \ \ \ : :
##                         IQR (CV) : 0 (0)                                        . \ \ \ \ . : : : :
##                                                                                 : : \ \ : : : : : :
##
## 24   misc_val\          Mean (sd) : 54.4 (590.4)\        26 distinct values     :\
##      [integer]          min < med < max:\                                       :\
##                         0 < 0 < 15500\                                          :\
##                         IQR (CV) : 0 (10.8)                                     :\
##                                                                                 :
##
## 25   year_sold\         Mean (sd) : 2007.9 (1.3)\        2006 : 259 (18.0%)\    III \
##      [integer]          min < med < max:\                2007 : 339 (23.5%)\    IIII \
##                         2006 < 2008 < 2010\              2008 : 323 (22.4%)\    IIII \
##                         IQR (CV) : 2 (0)                 2009 : 347 (24.1%)\    IIII \
##                                                          2010 : 172 (11.9%)     II
##
## 26   sale_price\        Mean (sd) : 177568.5 (73659.4)\  599 distinct values    \ \ :\
##      [integer]          min < med < max:\                                       \ \ :\
##                         52000 < 159000 < 755000\                                \ \ :\
##                         IQR (CV) : 77000 (0.4)                                  . : :\
##                                                                                 : : : .
## ---------------------------------------------------------------------------------------------------
```

Multiple linear regression

```
#fit1 <- lm(sale_price ~ )
```