IMT 573: Problem Set 6 - Regression

Jay Kuo

2022/11/16

Collaborators:

Instructions: Before beginning this assignment, please ensure you have access to R and RStudio.

1. Download the `problemset6.Rmd` file from Canvas. Open `problemset6.Rmd` in RStudio and supply your solutions to the assignment by editing `problemset6.Rmd`.
2. Replace the "Insert Your Name Here" text in the `author:` field with your own full name. Any collaborators must be listed on the top of your assignment.
3. All materials and resources that you use (with the exception of lecture slides) must be appropriately referenced within your assignment. In particular, note that Stack Overflow is licenses as Creative Commons (CC-BY-SA). This means you have to attribute any code you refer from SO.
4. Partial credit will be awarded for each question for which a serious attempt at finding an answer has been shown. But please **DO NOT** submit pages and pages of hard-to-read code and attempts that is impossible to grade. That is, avoid redundancy. Remember that one of the key goals of a data scientist is to produce coherent reports that others can easily follow. Students are *strongly* encouraged to attempt each question and to document their reasoning process even if they cannot find the correct answer. If you would like to include R code to show this process, but it does not run without errors you can do so with the `eval=FALSE` option as follows:

```
a + b # these object don't exist
# if you run this on its own it with give an error
```

6. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit PDF`, rename the knitted PDF file to `ps6_YourLastName_YourFirstName.pdf`, and submit the PDF file on Canvas.
7. Collaboration is often fun and useful, but each student must turn in an individual write-up in their own words as well as code/work that is their own. Regardless of whether you work with others, what you turn in must be your own work; this includes code and interpretation of results. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.

Setup In this problem set you will need, at minimum, the following R packages.

```
# Load standard libraries
library(tidyverse)
library(AmesHousing)
```

Housing Values in Ames, Iowa

In this problem we will use the Ames Housing dataset that is available as part of the AmesHousing package. This dataset contains information about home sales in the town of Ames, Iowa. Information on variable names and other details can be found in the AmesHousing package documentation as well as

here: http://jse.amstat.org/v19n3/decock/DataDocumentation.txt. Use this data to answer the following questions.

Question 1: Load the package and use the `make_ames()` to store the dataset. Describe what this function does.

Question

- Loads data and describes function.

Answer

```
?make_ames()
dt <- make_ames()
```

Question 2: Consider this data in context - what is the response variable of interest for a dataset on home sales? Filter the data to only contain observations where the Sale_Condition was "Normal." Select the following variables from the data and describe what each means: Lot_Frontage, Lot_Area, Bldg_Type, Overall_Qual, Overall_Cond, Year_Built, Gr_Liv_Area, TotRms_AbvGrd, Fireplaces, Garage_Cars, Garage_Area, Wood_Deck_SF, Total_Bsmt_SF, Full_Bath, Half_Bath, Year_Sold, and Sale_Price

Question

- Filters to Normal sales
- Names the correct variable of interest for the response variable.

Answer

```
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
##
##     set_names
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
dt %<>%
  filter(Sale_Condition == "Normal") %>%
  select(Lot_Frontage, Lot_Area, Bldg_Type, Overall_Qual, Overall_Cond, Year_Built, Gr_Liv_Area, TotRms
dt
```

```
## # A tibble: 2,413 x 17
##    Lot_Frontage Lot_Area Bldg_~1 Overa~2 Overa~3 Year_~4 Gr_Li~5 TotRm~6 Firep~7
##           <dbl>    <int> <fct>   <fct>   <fct>     <int>   <int>   <int>   <int>
##  1          141    31770 OneFam  Above_~ Average    1960    1656       7       2
##  2           80    11622 OneFam  Average Above_~    1961     896       5       0
##  3           81    14267 OneFam  Above_~ Above_~    1958    1329       6       0
##  4           93    11160 OneFam  Good    Average    1968    2110       8       2
##  5           74    13830 OneFam  Average Average    1997    1629       6       1
##  6           78     9978 OneFam  Above_~ Above_~    1998    1604       7       1
##  7           41     4920 TwnhsE  Very_G~ Average    2001    1338       6       0
##  8           43     5005 TwnhsE  Very_G~ Average    1992    1280       5       0
##  9           39     5389 TwnhsE  Very_G~ Average    1995    1616       5       1
```

```
## 10           60     7500 OneFam  Good    Average    1999    1804       7       1
## # ... with 2,403 more rows, 8 more variables: Garage_Cars <dbl>,
## #   Garage_Area <dbl>, Wood_Deck_SF <int>, Total_Bsmt_SF <dbl>,
## #   Full_Bath <int>, Half_Bath <int>, Year_Sold <int>, Sale_Price <int>, and
## #   abbreviated variable names 1: Bldg_Type, 2: Overall_Qual, 3: Overall_Cond,
## #   4: Year_Built, 5: Gr_Liv_Area, 6: TotRms_AbvGrd, 7: Fireplaces
```

Question

- Describes each variable

Answer

- Lot_Frontage: Linear feet of street connected to property
- Lot_Area: Lot size in square feet
- Bldg_Type: Type of dwelling (+)
- Overall_Qual: Rates the overall material and finish of the house (+)
- Overall_Cond: Rates the overall condition of the house (+)
- Year_Built: Original construction date (+)
- Gr_Liv_Area: Above grade (ground) living area square feet
- TotRms_AbvGrd: Total rooms above grade (does not include bathrooms)
- Fireplaces: Number of fireplaces
- Garage_Cars: Size of garage in car capacity
- Garage_Area: Size of garage in square feet
- Wood_Deck_SF: Wood deck area in square feet
- Total_Bsmt_SF: Total square feet of basement area
- Full_Bath: Full bathrooms above grade
- Half_Bath: Half baths above grade
- Year_Sold: Year Sold (YYYY) (+)
- Sale_Price: Sale price

Question 3: Provide a brief dive into the data and discuss any salient aspects of the variables: missingness, ranges, distributions, etc. Does each observation have complete data (Hint: you can use the complete.cases function in R)?

Question

- Describes salient aspects of the data.

Answer

```
# structure
str(data.frame(dt))
```

```
## 'data.frame':    2413 obs. of  17 variables:
##  $ Lot_Frontage : num  141 80 81 93 74 78 41 43 39 60 ...
##  $ Lot_Area     : int  31770 11622 14267 11160 13830 9978 4920 5005 5389 7500 ...
##  $ Bldg_Type    : Factor w/ 5 levels "OneFam","TwoFmCon",..: 1 1 1 1 1 1 5 5 5 1 ...
##  $ Overall_Qual : Factor w/ 10 levels "Very_Poor","Poor",..: 6 5 6 7 5 6 8 8 8 7 ...
##  $ Overall_Cond : Factor w/ 10 levels "Very_Poor","Poor",..: 5 6 6 5 5 6 5 5 5 5 ...
##  $ Year_Built   : int  1960 1961 1958 1968 1997 1998 2001 1992 1995 1999 ...
##  $ Gr_Liv_Area  : int  1656 896 1329 2110 1629 1604 1338 1280 1616 1804 ...
##  $ TotRms_AbvGrd: int  7 5 6 8 6 7 6 5 5 7 ...
##  $ Fireplaces   : int  2 0 0 2 1 1 0 0 1 1 ...
##  $ Garage_Cars  : num  2 1 1 2 2 2 2 2 2 2 ...
##  $ Garage_Area  : num  528 730 312 522 482 470 582 506 608 442 ...
##  $ Wood_Deck_SF : int  210 140 393 0 212 360 0 0 237 140 ...
##  $ Total_Bsmt_SF: num  1080 882 1329 2110 928 ...
```

```
##  $ Full_Bath    : int  1 1 1 2 2 2 2 2 2 2 ...
##  $ Half_Bath    : int  0 0 1 1 1 1 0 0 0 1 ...
##  $ Year_Sold    : int  2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 ...
##  $ Sale_Price   : int  215000 105000 172000 244000 189900 195500 213500 191500 236500 189000 ...
```

```
# summary statistics
summary(dt)
```

```
##   Lot_Frontage       Lot_Area         Bldg_Type          Overall_Qual
##  Min.   :  0.00   Min.   :  1300   OneFam  :2002   Average      :715
##  1st Qu.: 37.00   1st Qu.:  7390   TwoFmCon:  52   Above_Average:640
##  Median : 60.00   Median :  9360   Duplex  :  78   Good         :493
##  Mean   : 55.46   Mean   : 10060   Twnhs   :  93   Very_Good    :256
##  3rd Qu.: 77.00   3rd Qu.: 11404   TwnhsE  : 188   Below_Average:185
##  Max.   :313.00   Max.   :215245                   Excellent    : 64
##                                                    (Other)      : 60
##        Overall_Cond    Year_Built    Gr_Liv_Area   TotRms_AbvGrd
##  Average      :1282   Min.   :1872   Min.   : 334   Min.   : 2.000
##  Above_Average: 474   1st Qu.:1953   1st Qu.:1100   1st Qu.: 5.000
##  Good         : 352   Median :1971   Median :1431   Median : 6.000
##  Very_Good    : 139   Mean   :1969   Mean   :1477   Mean   : 6.367
##  Below_Average:  82   3rd Qu.:1998   3rd Qu.:1724   3rd Qu.: 7.000
##  Excellent    :  39   Max.   :2010   Max.   :4316   Max.   :13.000
##  (Other)      :  45
##    Fireplaces     Garage_Cars     Garage_Area      Wood_Deck_SF
##  Min.   :0.000   Min.   :0.000   Min.   :   0.0   Min.   :   0.00
##  1st Qu.:0.000   1st Qu.:1.000   1st Qu.: 315.0   1st Qu.:   0.00
##  Median :1.000   Median :2.000   Median : 472.0   Median :   0.00
##  Mean   :0.603   Mean   :1.732   Mean   : 461.3   Mean   :  95.87
##  3rd Qu.:1.000   3rd Qu.:2.000   3rd Qu.: 576.0   3rd Qu.: 168.00
##  Max.   :4.000   Max.   :5.000   Max.   :1488.0   Max.   :1424.00
##
##  Total_Bsmt_SF    Full_Bath       Half_Bath       Year_Sold      Sale_Price
##  Min.   :   0   Min.   :0.000   Min.   :0.000   Min.   :2006   Min.   : 35000
##  1st Qu.: 784   1st Qu.:1.000   1st Qu.:0.000   1st Qu.:2007   1st Qu.:129500
##  Median : 970   Median :2.000   Median :0.000   Median :2008   Median :159000
##  Mean   :1023   Mean   :1.539   Mean   :0.378   Mean   :2008   Mean   :175568
##  3rd Qu.:1246   3rd Qu.:2.000   3rd Qu.:1.000   3rd Qu.:2009   3rd Qu.:206900
##  Max.   :3206   Max.   :4.000   Max.   :2.000   Max.   :2010   Max.   :755000
##
```

```
# check missing value
sum(complete.cases(dt) != 1)
```

```
## [1] 0
```

There are 0 missing value.

Question 4: For each predictor, fit a simple (i.e. using only the one variable) linear regression model to predict the home sale price. Dummify variables as/when needed. In which of the models is there a statistically significant association between the predictor and the response? Describe your results.

Question

- Fits regression model for each variable.

Answer

```
dt_new <- dt

# assign `Bldg_Type = 1` if Bldg_Type == "OneFam"
dt_new$Bldg_Type <- ifelse(dt$Bldg_Type == "OneFam", 1, 0)

# Assign `Overall_Qual = 1` if Overall_Qual > Average
dt_new$Overall_Qual <- ifelse(dt$Overall_Qual %in% c("Above_Average", "Good", "Very_Good", "Excellent",

# Assign `Overall_Cond = 1` if Overall_Cond > Average
dt_new$Overall_Cond <- ifelse(dt$Overall_Cond %in% c("Above_Average", "Good", "Very_Good", "Excellent",

# Assign `Year_Built = 1` if Year_Built >= 1971
dt_new$Year_Built <- ifelse(dt$Year_Built >= 1971, 1, 0)


# Assign `Year_Sold = 1` if Year_Sold >= 2008
dt_new$Year_Sold <- ifelse(dt$Year_Sold >= 2008, 1, 0)

dt_new
```

```
## # A tibble: 2,413 x 17
##    Lot_Frontage Lot_Area Bldg_~1 Overa~2 Overa~3 Year_~4 Gr_Li~5 TotRm~6 Firep~7
##           <dbl>    <int>   <dbl>   <dbl>   <dbl>   <dbl>   <int>   <int>   <int>
##  1          141    31770       1       1       0       0    1656       7       2
##  2           80    11622       1       0       1       0     896       5       0
##  3           81    14267       1       1       1       0    1329       6       0
##  4           93    11160       1       1       0       0    2110       8       2
##  5           74    13830       1       0       0       1    1629       6       1
##  6           78     9978       1       1       1       1    1604       7       1
##  7           41     4920       0       1       0       1    1338       6       0
##  8           43     5005       0       1       0       1    1280       5       0
##  9           39     5389       0       1       0       1    1616       5       1
## 10           60     7500       1       1       0       1    1804       7       1
## # ... with 2,403 more rows, 8 more variables: Garage_Cars <dbl>,
## #   Garage_Area <dbl>, Wood_Deck_SF <int>, Total_Bsmt_SF <dbl>,
## #   Full_Bath <int>, Half_Bath <int>, Year_Sold <dbl>, Sale_Price <int>, and
## #   abbreviated variable names 1: Bldg_Type, 2: Overall_Qual, 3: Overall_Cond,
## #   4: Year_Built, 5: Gr_Liv_Area, 6: TotRms_AbvGrd, 7: Fireplaces
```

```
SLR <- list()

y <- c("Sale_Price")
controls <- colnames(dt)[-17]

for (i in 1:16){
  #print(paste("Start the iteration: ", i))
  SLR[[i]] <- lm(formula= as.formula(paste(y, controls[i], sep = " ~ ")), data = dt_new)
}
```

```
library(jtools); library(huxtable)
```

```
##
## Attaching package: 'huxtable'
```

```
## The following object is masked from 'package:dplyr':
##
##     add_rownames
## The following object is masked from 'package:ggplot2':
##
##     theme_grey
```

```
export_summs(SLR[[1]], SLR[[2]], SLR[[3]], SLR[[4]])
```

	Model 1	Model 2	Model 3	Model 4
(Intercept)	156742.56 ***	151680.76 ***	157949.15 ***	127982.52 ***
	(2757.04)	(2195.89)	(3479.67)	(1946.74)
Lot_Frontage	339.42 ***			
	(42.54)			
Lot_Area		2.37 ***		
		(0.17)		
Bldg_Type			21235.48 ***	
			(3820.18)	
Overall_Qual				78217.24 ***
				(2495.88)
N	2413	2413	2413	2413
R2	0.03	0.08	0.01	0.29

*** p < 0.001; ** p < 0.01; * p < 0.05.

```
export_summs(SLR[[5]], SLR[[6]], SLR[[7]], SLR[[8]])
```

```
export_summs(SLR[[9]], SLR[[10]], SLR[[11]], SLR[[12]])
```

```
export_summs(SLR[[13]], SLR[[14]], SLR[[15]], SLR[[16]])
```

Question

- Correctly answers where there is a statistically significant association.
- Describes results.

Answer

- Excluding `Year Sold`, other covariates are significant individually at 5% level.

Question 5: Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis $H_0 : \beta_j = 0$?

Question

- Fits a multiple regression model to predict the response using all of the predictors.

	Model 1	Model 2	Model 3	Model 4
(Intercept)	190420.81 ***	139689.74 ***	15784.33 ***	27866.72 ***
	(1832.27)	(1790.04)	(3148.71)	(5406.40)
Overall_Cond	-35697.89 ***			
	(2840.55)			
Year_Built		70499.49 ***		
		(2509.24)		
Gr_Liv_Area			108.15 ***	
			(2.03)	
TotRms_AbvGrd				23198.75 ***
				(825.98)
N	2413	2413	2413	2413
R2	0.06	0.25	0.54	0.25

*** p < 0.001; ** p < 0.01; * p < 0.05.

Answer

```
# y ~ x1 + x2 ...
MLR<- dt_new %>%
  lm(formula = as.formula(paste(y, paste(c(controls), collapse = "+"), sep = " ~ ")), data = .)

summary(MLR)
```

```
##
## Call:
## lm(formula = as.formula(paste(y, paste(c(controls), collapse = "+"),
##     sep = " ~ ")), data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -135839  -17694   -1101   15803  250523
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.256e+04  3.659e+03  -6.166 8.19e-10 ***
## Lot_Frontage   1.132e+02  2.032e+01   5.569 2.85e-08 ***
## Lot_Area       4.352e-01  8.511e-02   5.113 3.41e-07 ***
## Bldg_Type      1.223e+04  1.869e+03   6.542 7.39e-11 ***
## Overall_Qual   1.187e+04  1.734e+03   6.842 9.89e-12 ***
## Overall_Cond   4.636e+03  1.495e+03   3.101  0.00195 **
## Year_Built     2.879e+04  1.944e+03  14.807  < 2e-16 ***
## Gr_Liv_Area    7.398e+01  3.116e+00  23.745  < 2e-16 ***
## TotRms_AbvGrd -5.461e+03  7.531e+02  -7.251 5.57e-13 ***
```

	Model 1	Model 2	Model 3	Model 4
(Intercept)	143026.19 ***	68827.95 ***	74609.67 ***	157274.15 ***
	(1716.22)	(2888.37)	(2825.60)	(1681.12)
Fireplaces	53967.37 ***			
	(1937.65)			
Garage_Cars		61632.66 ***		
		(1537.40)		
Garage_Area			218.88 ***	
			(5.61)	
Wood_Deck_SF				190.81 ***
				(10.40)
N	2413	2413	2413	2413
R2	0.24	0.40	0.39	0.12

*** p < 0.001; ** p < 0.01; * p < 0.05.

```
## Fireplaces     8.948e+03  1.188e+03   7.532 7.06e-14 ***
## Garage_Cars    5.276e+03  2.087e+03   2.528  0.01154 *
## Garage_Area    3.533e+01  7.226e+00   4.889 1.08e-06 ***
## Wood_Deck_SF   2.535e+01  5.384e+00   4.708 2.65e-06 ***
## Total_Bsmt_SF  4.910e+01  2.021e+00  24.298  < 2e-16 ***
## Full_Bath     -3.571e+03  1.889e+03  -1.891  0.05873 .
## Half_Bath      1.576e+03  1.697e+03   0.929  0.35322
## Year_Sold      7.093e+02  1.309e+03   0.542  0.58804
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31660 on 2396 degrees of freedom
## Multiple R-squared:  0.8023, Adjusted R-squared:  0.801
## F-statistic: 607.9 on 16 and 2396 DF,  p-value: < 2.2e-16
```

Question

- Correctly answers which predictors allow for rejecting the null hypothesis.

Answer

- Excluding `Full_Bath`, `Half_Bath`, and `Year_Sold`, other covariates are significant individually at 5% level.

Question 6: How do your results from (4) compare to your results from (5)? You need to compare the coefficients across the two models and report on the changes you observe and reasons why. What happened to the coefficients? What happened to the p-values? Why?

Question

	Model 1	Model 2	Model 3	Model 4
(Intercept)	61539.79 ***	65522.56 ***	159835.64 ***	173310.84 ***
	(2983.95)	(3622.74)	(1734.65)	(2224.28)
Total_Bsmt_SF	111.48 ***			
	(2.71)			
Full_Bath		71496.71 ***		
		(2218.92)		
Half_Bath			41624.26 ***	
			(2773.34)	
Year_Sold				3903.70
				(2925.37)
N	2413	2413	2413	2413
R2	0.41	0.30	0.09	0.00

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

- Compares results from two questions.

Answer

- Compared to model (4), `Full_Bath` and `Half_Bath` are insignificant in model (5) at 5% level.

Question

- Explains what happens to p-values and coefficients.

Answer

- 1. Absolute values of coefficients in MLR are smaller than in SLR.
- 2. Most p-value of covariates in MLR are smaller than in SLR.

```r
# SLR results
SLR_coef <- c()
SLR_pval <- c()
for (i in 1:16){
  SLR_coef[i] <- summary(SLR[[i]])$coefficients[2, 1]
  SLR_pval[i] <- summary(SLR[[i]])$coefficients[2, 4]
}

# MLR results
MLR_coef <- summary(MLR)$coefficients[2:17, 1]
MLR_pval <- summary(MLR)$coefficients[2:17, 4]

# comparison
cbind(SLR_coef, MLR_coef)
```

```
##                    SLR_coef      MLR_coef
## Lot_Frontage     339.420111   113.1594461
## Lot_Area           2.374392     0.4352239
## Bldg_Type      21235.478455 12229.5538070
## Overall_Qual   78217.236554 11865.5172263
## Overall_Cond  -35697.892613  4636.3426507
## Year_Built     70499.491941 28790.4154928
## Gr_Liv_Area      108.150940    73.9823262
## TotRms_AbvGrd  23198.745727 -5460.6163740
## Fireplaces     53967.367265  8947.8343056
## Garage_Cars    61632.660758  5276.2642807
## Garage_Area      218.877164    35.3333027
## Wood_Deck_SF     190.806806    25.3456374
## Total_Bsmt_SF    111.482963    49.1040730
## Full_Bath      71496.710116 -3571.3717078
## Half_Bath      41624.260708  1576.1091483
## Year_Sold       3903.701788   709.2732210
```

```
cbind(SLR_pval, MLR_pval)
```

```
##                    SLR_pval      MLR_pval
## Lot_Frontage   2.256004e-15  2.846996e-08
## Lot_Area       3.790598e-43  3.413213e-07
## Bldg_Type      3.016114e-08  7.389004e-11
## Overall_Qual  3.829400e-181  9.890253e-12
## Overall_Cond   3.934324e-35  1.953709e-03
## Year_Built    1.704507e-150  1.562281e-47
## Gr_Liv_Area    0.000000e+00 4.166410e-112
## TotRms_AbvGrd 2.095108e-150  5.571813e-13
## Fireplaces    2.971400e-148  7.057584e-14
## Garage_Cars   9.993043e-270  1.153544e-02
## Garage_Area   2.545845e-258  1.078276e-06
## Wood_Deck_SF   1.572202e-70  2.647692e-06
## Total_Bsmt_SF 6.849081e-281 9.169564e-117
## Full_Bath     9.729183e-190  5.873080e-02
## Half_Bath      9.554834e-49  3.532217e-01
## Year_Sold      1.821892e-01  5.880359e-01
```