CSI 2300: Intro to Data Science

In-Class Exercise 20: Modeling – Multiple Linear Regression

For this lecture, we'll return to another dataset we've used before; the Boulder Housing dataset[1].

1. Load the 2020 Boulder housing dataset in (`boulder-2020-residential_sales.csv`). Create several new numeric variables using the code given below. What are these new variables?

```
sales2020 <- read.csv(file="boulder-2020-residential_sales.csv",
                      header=T, stringsAsFactors=F)

# Remove $ and , (dollar signs and commas) from everywhere, then parse as
# integer. In the pattern, the vertical bar "|" means "or".
to_remove_pattern <- '\\$|,'
sales2020$BLDG_VALUE <- as.integer(gsub(to_remove_pattern, '', sales2020$BLDG_VALUE))
sales2020$LAND_VALUE <- as.integer(gsub(to_remove_pattern, '', sales2020$LAND_VALUE))
sales2020$SALE_PRICE <- as.integer(gsub(to_remove_pattern, '', sales2020$SALE_PRICE))

# add together all baths
sales2020$total_baths = sales2020$FULL_BATHS +
                        sales2020$THREE_QTR_BATHS * 0.75 +
                        sales2020$HALF_BATHS * 0.5

# add together all square footage
sales2020$total_sqft = sales2020$STUDIO_SQFT +
                       sales2020$GARAGE_SQFT +
                       sales2020$ABOVE_GROUND_SQFT +
                       sales2020$FINISHED_BSMT_SQFT +
                       sales2020$FINISHED_GARAGE_SQFT +
                       sales2020$UNFINISHED_BSMT_SQFT
```

2. Build a multiple linear regression model to predict the `SALE_PRICE` (dependent variable) using the following independent variables:

   - `BLDG_VALUE`,
   - `LAND_VALUE`,
   - `total_baths`,
   - `total_sqft`, and
   - `BLDG1_YEAR_BUILT`

[1] https://www.bouldercounty.org/property-and-land/assessor/sales/recent/

```r
# Build the model Error in eval(mf, parent.frame()) : object 'sales2020' not found
model <- lm(SALE_PRICE ~ BLDG_VALUE + LAND_VALUE + total_baths + total_sqft + BLDG1_YEAR
            data = sales2020)
summary(model)
#
# Call:
# lm(formula = SALE_PRICE ~ BLDG_VALUE + LAND_VALUE + total_baths +
#     total_sqft + BLDG1_YEAR_BUILT, data = sales2020)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -1297159   -86517   -27672    46180  3655080
#
# Coefficients:
#                     Estimate Std. Error t value Pr(>|t|)
# (Intercept)      -3.241e+06  3.724e+05  -8.703  < 2e-16 ***
# BLDG_VALUE        7.868e-01  1.649e-02  47.718  < 2e-16 ***
# LAND_VALUE        1.221e+00  2.109e-02  57.914  < 2e-16 ***
# total_baths       1.219e+04  6.398e+03   1.906   0.0568 .
# total_sqft        2.281e+01  3.839e+00   5.942 3.14e-09 ***
# BLDG1_YEAR_BUILT  1.656e+03  1.896e+02   8.733  < 2e-16 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 206600 on 3046 degrees of freedom
# Multiple R-squared:  0.7857,  Adjusted R-squared:  0.7853
# F-statistic:  2233 on 5 and 3046 DF,  p-value: < 2.2e-16
```

What do your coefficients look like? Do they make sense? When discussing this, consider both the coefficients' values and their signs.

The intercept is negative, which doesn't make sense in the context of this problem. This is because the model is not valid for houses with zero value.

3. Using the model you just built, construct some predictions of prices for the following (theoretical) houses:

House	BLDG_VALUE	LAND_VALUE	total_baths	total_sqft	BLDG1_YEAR_BUILT
A	400000	100000	3	3000	1995
B	200000	50000	1	1500	2005
C	2000000	3000000	9	20000	1800

First, create a new data frame using these values of the new homes. Be sure to use the exact same variable names for each of the columns that are used in the original data frame.

Then, use the `predict` function with the arguments as follows: `predict(your_model, new_dataframe)`

```
# Create a new data frame with the theoretical houses
theoretical_houses <- data.frame(
  BLDG_VALUE = c(400000, 200000, 2000000),
  LAND_VALUE = c(100000, 50000, 3000000),
  total_baths = c(3, 1, 9),
  total_sqft = c(3000, 1500, 20000),
  BLDG1_YEAR_BUILT = c(1995, 2005, 1800)
)

predicted_prices <- predict(model, newdata = theoretical_houses)
predicted_prices
#         1         2         3
#  603497.0  343041.9 5541901.9
```

4. Though we can use our linear model to make predictions for any type of house with (e.g.) any number of baths, such predictions don't always make sense. For which of the theoretical houses listed in the previous question does it make sense for our model to make a prediction, and why?

House A and B have values within the range of the original data, so predictions are likely reasonable. House C has extremely high value which is outside the range of the training data, so the prediction may not be reliable.

5. Consider all the variables we have looked at for this problem (independent and dependent). Find the correlation between each of the independent variables and the dependent variable. Then find the correlation between each pair of independent variables. Which do you find are the most strongly correlated (close to +/-1) and least strongly correlated (close to 0)?

```
vars <- sales2020[, c("SALE_PRICE", "BLDG_VALUE", "LAND_VALUE", "total_baths", "total_sq

cor_matrix <- cor(vars, use = "complete.obs")
cor_matrix
#                 SALE_PRICE BLDG_VALUE LAND_VALUE total_baths total_sqft
# SALE_PRICE        1.0000000  0.6731334  0.7219106   0.4729272  0.4977687
# BLDG_VALUE        0.6731334  1.0000000  0.2780071   0.4162028  0.4103543
# LAND_VALUE        0.7219106  0.2780071  1.0000000   0.2730219  0.3023440
# total_baths       0.4729272  0.4162028  0.2730219   1.0000000  0.7536025
# total_sqft        0.4977687  0.4103543  0.3023440   0.7536025  1.0000000
# BLDG1_YEAR_BUILT  0.0211615  0.1377111 -0.2930836   0.4071029  0.3740742
```

```
#                    BLDG1_YEAR_BUILT
# SALE_PRICE               0.0211615
# BLDG_VALUE               0.1377111
# LAND_VALUE              -0.2930836
# total_baths              0.4071029
# total_sqft               0.3740742
# BLDG1_YEAR_BUILT         1.0000000
```

The variable most strongly correlated with sale price is land value, followed by bldg value. Among the independent variables, total_baths and total_sqft are highly correlated with each other, which could indicate potential correlated if both are included in a model.

6. "Occam's Razor"[2] says that simpler theories are preferable to complex ones, if both have the ability to explain. When used in modeling, this principle leads us to prefer simple models (using fewer variables) that are still useful (i.e., still make reasonable predictions).

```
simple_model <- lm(SALE_PRICE ~ BLDG_VALUE + total_sqft, data = sales2020)
summary(simple_model)
#
# Call:
# lm(formula = SALE_PRICE ~ BLDG_VALUE + total_sqft, data = sales2020)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -1196577  -165703   -70749    88285  4654276
#
# Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
# (Intercept) 6.095e+04  1.290e+04   4.724 2.41e-06 ***
# BLDG_VALUE  9.800e-01  2.411e-02  40.646  < 2e-16 ***
# total_sqft  7.682e+01  4.000e+00  19.206  < 2e-16 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 311600 on 3049 degrees of freedom
# Multiple R-squared:  0.5121,  Adjusted R-squared:  0.5118
# F-statistic:  1600 on 2 and 3049 DF,  p-value: < 2.2e-16
```

With Occam's Razor in mind, using the results from the previous question, choose a *subset* of the independent variables that are *most* correlated with the dependent variable and *least* correlated with each other. Use them to make a new linear model. Compare the coefficients

[2] https://en.wikipedia.org/wiki/Occam%27s_razor

and fit (e.g. adjusted R^2) for this model with the previous model that used more independent variables.

The simpler model using only bldg value and total sqft explains the variance in sail price (adjusted R-squared = 0.5118), with both predictors highly significant. Compared to the full model, this simpler model is easier to interpret.

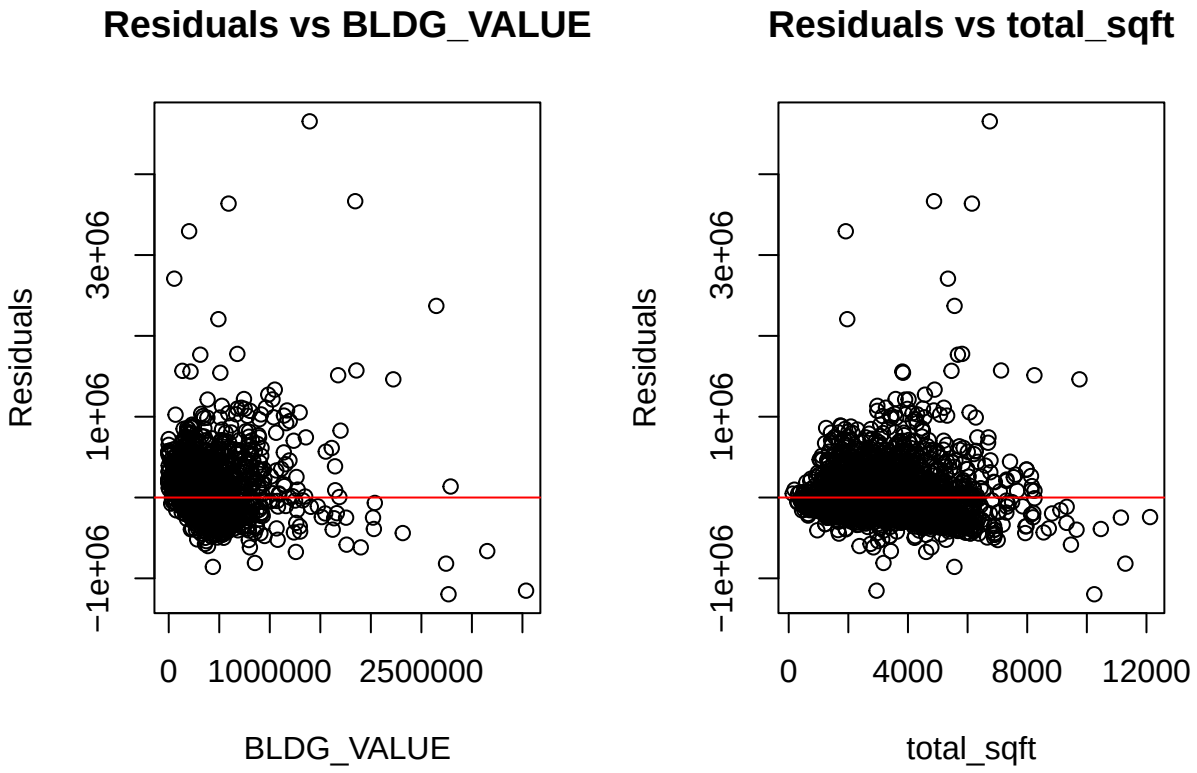
7. Calculate the residuals, and for each of the independent variables you used in your most recent model, make a plot of the residuals (y-axis) versus the independent variable (x-axis). Look carefully. Do there appear to be any patterns between the residuals and any of the independent variables? If there are any patterns, what might that tell you about the data?

```r
sales2020$simple_resid <- residuals(simple_model)

par(mfrow = c(1, 2)) # 2 plots side by side

plot(
  sales2020$BLDG_VALUE, sales2020$simple_resid,
  xlab = "BLDG_VALUE", ylab = "Residuals",
  main = "Residuals vs BLDG_VALUE"
)
abline(h = 0, col = "red")

plot(
  sales2020$total_sqft, sales2020$simple_resid,
  xlab = "total_sqft", ylab = "Residuals",
  main = "Residuals vs total_sqft"
)
abline(h = 0, col = "red")
```



There do not appear to be strong patterns the points are fairly randomly scattered around zero. However, there is some spread and a few large outliers, especially at higher values, which suggests that the model may not fit extremely high-value. This indicates that the relationship between the predictors and sale price is not linear for all ranges.