Exercise 12: Housing Data

David Brehm

2021-01-31

a. Explain why you chose to remove data points from your 'clean' dataset.

I chose to keep Sale Price, Building Grade, Square Feet Total Living, Bedrooms, Square Feet Lot, and Bathrooms. "Bathrooms" is combined full, half, and three-quarters bathroom columns. Most of the other columns should not have an effect on our independent variable Sale Price. Property Type and Postal Location only had one value as well.

b. Create two models. One that will contain the variables Sale Price and Square Foot of Lot (same variables used from previous assignment on simple regression) and one that will contain Sale Price and several additional predictors of your choice.

```
mod1 <- lm(sale_price ~ sq_ft_lot, data=data)
mod2 <- lm(sale_price ~ building_grade + square_feet_total_living + bedrooms + sq_ft_lot + bath, data=da
```

c. Execute a summary() function on two variables defined in the previous step to compare the model results.

```
##
## Call:
## lm(formula = sale_price ~ sq_ft_lot, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2016064  -194842   -63293    91565  3735109
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.418e+05  3.800e+03  168.90   <2e-16 ***
## sq_ft_lot   8.510e-01  6.217e-02   13.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 401500 on 12863 degrees of freedom
## Multiple R-squared:  0.01435,    Adjusted R-squared:  0.01428
## F-statistic: 187.3 on 1 and 12863 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = sale_price ~ building_grade + square_feet_total_living +
##     bedrooms + sq_ft_lot + bath, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1962684  -113545   -41956    40076  3754813
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -3.968e+04  3.170e+04  -1.252 0.210641
## building_grade            4.005e+04  4.422e+03   9.056  < 2e-16 ***
## square_feet_total_living  1.487e+02  6.465e+00  22.993  < 2e-16 ***
## bedrooms                 -2.000e+04  4.564e+03  -4.382 1.19e-05 ***
## sq_ft_lot                 1.279e-01  5.794e-02   2.207 0.027309 *
## bath                      2.408e+04  6.975e+03   3.452 0.000558 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 358400 on 12859 degrees of freedom
## Multiple R-squared:  0.2148, Adjusted R-squared:  0.2145
## F-statistic: 703.6 on 5 and 12859 DF,  p-value: < 2.2e-16
```

Model Variables	R2	Adjusted R2
Square Foot of Lot	0.01435	0.01428
Multiple	0.2148	0.2145

d. What are the standardized betas for each parameter and what do the values indicate?

```
##           building_grade square_feet_total_living                 bedrooms
##               0.10821336               0.36386458              -0.04332910
##                sq_ft_lot                     bath
##               0.01800524               0.04139706
```

Standardized beta values are measured in standard deviation units, so these values tell us the number of standard deviations that the outcome will change after changing a predictor by one standard deviation.

e. Calculate the confidence intervals for the parameters in your model and explain what the results indicate.

```
##                                2.5 %        97.5 %
## (Intercept)              -1.018186e+05  2.245201e+04
## building_grade            3.138161e+04  4.871813e+04
## square_feet_total_living  1.359808e+02  1.613265e+02
## bedrooms                 -2.894548e+04 -1.105206e+04
## sq_ft_lot                 1.432041e-02  2.414518e-01
## bath                      1.040719e+04  3.775283e+04
```

These confidence intervals indicate that there is a 95% probability that the true values of the coefficient are within that interval.

f. Assess the improvement of the new model compared to your original model (simple regression model) by testing whether this change is significant by performing an analysis of variance.

```
## Analysis of Variance Table
##
## Model 1: sale_price ~ sq_ft_lot
## Model 2: sale_price ~ building_grade + square_feet_total_living + bedrooms +
##     sq_ft_lot + bath
##   Res.Df        RSS Df  Sum of Sq     F    Pr(>F)
## 1  12863 2.0734e+15
## 2  12859 1.6517e+15  4 4.2167e+14 820.7 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

P-value of less than 0.05 indicates an improvement between models and we can reject the null hypothesis that the one-predictor model is as good as the multiple-predictor model.

g. Perform casewise diagnostics to identify outliers and/or influential cases, storing each function's output in a dataframe assigned to a unique variable name.

```
##      rstudent unadjusted p-value Bonferroni p
## 6438 9.334760         1.1763e-20   1.5134e-16
## 6437 9.334494         1.1793e-20   1.5171e-16
## 6441 9.334316         1.1813e-20   1.5197e-16
## 6433 9.334031         1.1844e-20   1.5237e-16
## 6434 9.333823         1.1867e-20   1.5267e-16
## 6430 9.333677         1.1884e-20   1.5288e-16
## 6442 9.332473         1.2018e-20   1.5462e-16
## 6439 9.331469         1.2132e-20   1.5608e-16
## 6431 9.331388         1.2141e-20   1.5620e-16
## 6429 9.329466         1.2362e-20   1.5904e-16
```

```
##       rstudent unadjusted p-value Bonferroni p
## 4649  10.57346         5.0591e-26   6.5086e-22
## 11992 10.53691         7.4415e-26   9.5735e-22
## 6430  10.52200         8.7067e-26   1.1201e-21
## 6437  10.48011         1.3520e-25   1.7393e-21
## 6438  10.46986         1.5052e-25   1.9365e-21
## 6431  10.39563         3.2669e-25   4.2028e-21
## 6436  10.36774         4.3649e-25   5.6154e-21
## 6432  10.26847         1.2168e-24   1.5654e-20
## 6433  10.26524         1.2579e-24   1.6182e-20
## 6434  10.26521         1.2583e-24   1.6188e-20
```

h. Calculate the standardized residuals using the appropriate command, specifying those that are +-2, storing the results of large residuals in a variable you create.

```
data$stdRes <- rstandard(mod2)
data$studRes <- rstudent(mod2)
data$largeResid <-  data$stdRes > 2 | data$stdRes < -2
```

i. Use the appropriate function to show the sum of large residuals.

```
sum(data$largeResid)
```

```
## [1] 319
```

j. Which specific variables have large residuals (only cases that evaluate as TRUE)?

```
## # A tibble: 319 x 9
##    building_grade square_feet_tot~ bedrooms sq_ft_lot  bath sale_price stdRes
##             <dbl>            <dbl>    <dbl>     <dbl> <dbl>      <dbl>  <dbl>
##  1             10             4920        4    112650  4.5      265000  -2.43
##  2              6              660        0    225640  1       1390000   2.90
##  3             10             3840        0    236966  0        229000  -2.05
##  4             11             5800        5     63162  4.5      390000  -2.48
##  5              9             3360        2      8752  2.5     1588359   2.08
##  6              6              900        2     14043  1       1450000   3.15
##  7              9             4710        4     18498  4        163000  -2.45
##  8             11             5060        4     89734 23.5      270000  -4.18
##  9             10             6880        5    288367  4.5      200000  -3.43
## 10             11             4490        4     55303  3.25     300000  -2.16
## # ... with 309 more rows, and 2 more variables: studRes <dbl>, largeResid <lgl>
```

k. Investigate further by calculating the leverage, cooks distance, and covariance rations. Comment on all cases that are problematics.

```
##       lev              cooksDist            covRatio
##  Min.   :0.0001273   Min.   :0.0001213   Min.   :0.9504
##  1st Qu.:0.0002718   1st Qu.:0.0010341   1st Qu.:0.9746
##  Median :0.0005273   Median :0.0020179   Median :0.9926
##  Mean   :0.0023849   Mean   :0.0100421   Mean   :0.9877
##  3rd Qu.:0.0010592   3rd Qu.:0.0040426   3rd Qu.:0.9987
##  Max.   :0.1489852   Max.   :0.8508115   Max.   :1.1661
```

None of these values have a Cooks Distance greater than 1, so none of these cases have an undue influence on the model.

l. Perform the necessary calculations to assess the assumption of independence and state if the condition is met or not.

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.7409137     0.5181686       0
##  Alternative hypothesis: rho != 0
```

The Durbin Watson test returns a value between 0 and 4. A result closer to 0 indicates a positive autocorrelation while a result closer to 4 indicates a negative autocorrelation. The closer to 2 the better. The test for this model returns 0.518, which is outside of the conservative range of 1 to 3 and therefore does not indicate independence.

m. Perform the necessary calculations to assess the assumption of no multicollinearity and state if the condition is met or not.

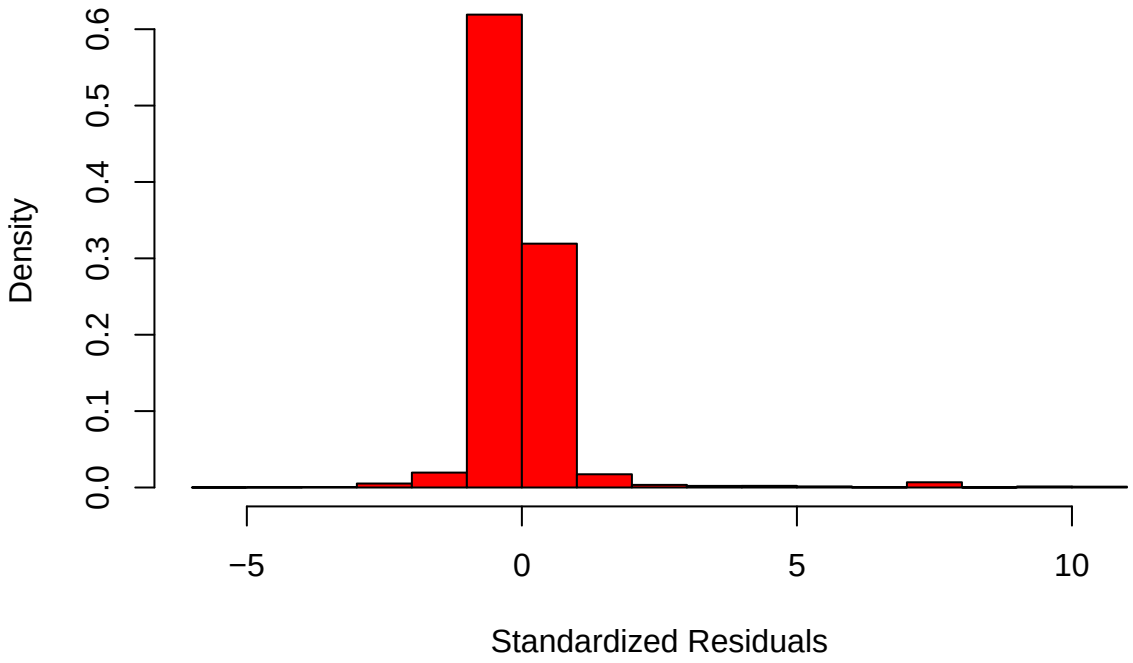
```
##           building_grade square_feet_total_living                 bedrooms
##                 2.338171                 4.101387                 1.601521
##                sq_ft_lot                     bath
##                 1.089678                 2.355022
```

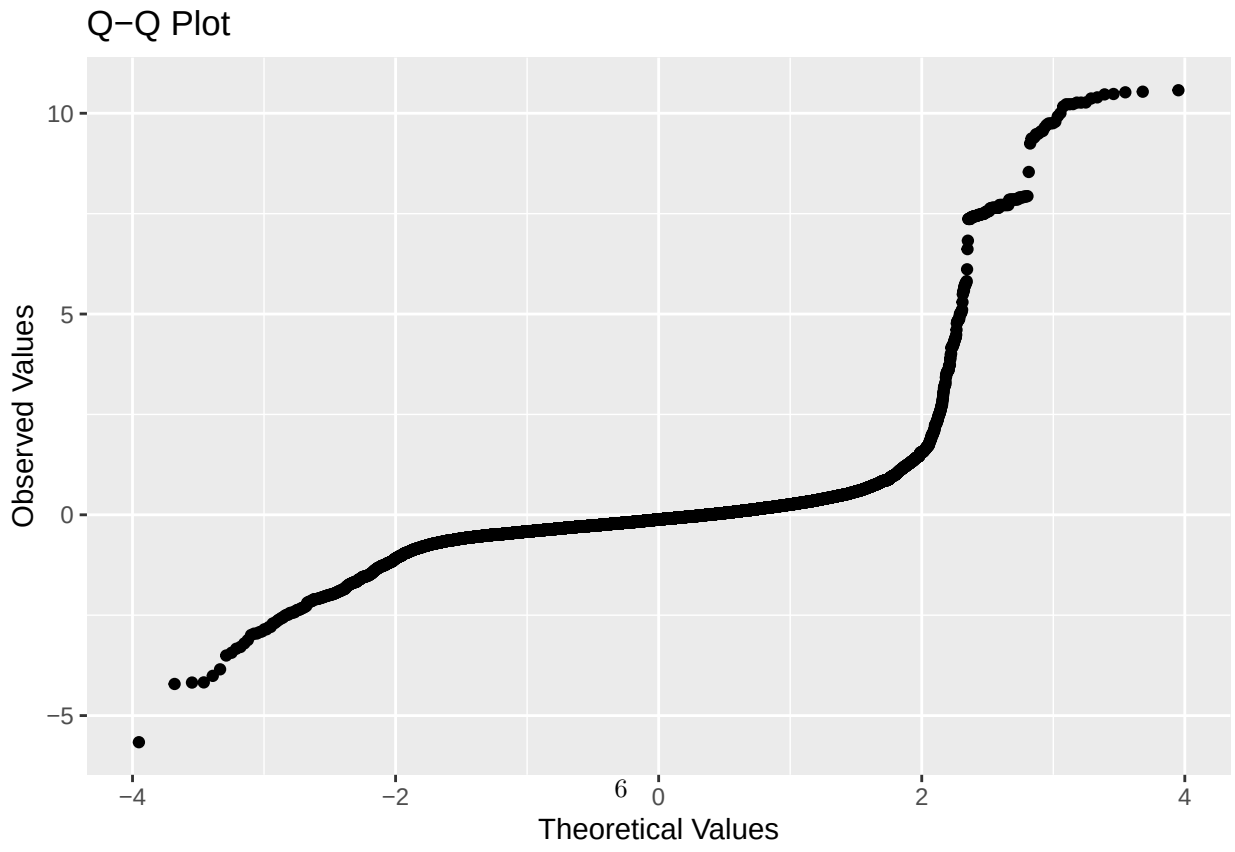
```
##           building_grade square_feet_total_living                 bedrooms
##                0.4276847                0.2438200                0.6244065
##                sq_ft_lot                     bath
##                0.9177019                0.4246246
```

```
## [1] 2.297156
```

The largest VIF is well below 10, there is no tolerance below 0.2, and the average VIF is not substantially greater than one. The condition of no multicollinearity is met.

n. Visually check the assumptions related to the residuals using the plot() and hist() functions. Summarize what each graph is informing you of and if any anomalies are present.





The Q-Q plot is not normal, it appears almost bimodal. The residuals are closely distributed around 0 though.

o. Overall, is this regression model unbiased? If an unbiased regression model, what does this tell us about the sample vs. the entire population model?

From the residual and VIF analysis, the model does not appear biased. This tells us that on average the sample regression model is the same as the entire population model. It is not a guarantee though.