PSTAT 126 Final Project

Data: Real Estate

Chloe Lam (Friday 9am)
Fady Naeim (Friday 12pm)
Shireen Mann (Friday 10am)
Caitlyn Jue (Thursday 9am)

Introduction

The real estate data was assembled by a city tax assessor, and comprises the data of 521 transactions in the residential home sale industry in a midwestern city during the year of 2002. It keeps a record of 12 variables in terms of price, size, architectural soundness, quality, and amenities. We will take into consideration some of these variables in terms of how they affect the sales price of residences.

Question of Interest

Do the possible predictor variables pertaining to size of the house, specifically the number of beds and baths, square feet, lot size, garage size, and the presence of a pool, or lack thereof, affect the sales price of the homes?

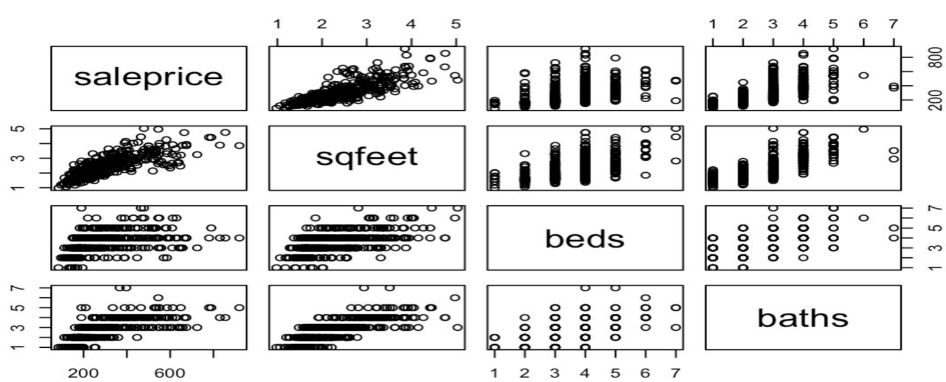
Regression Method

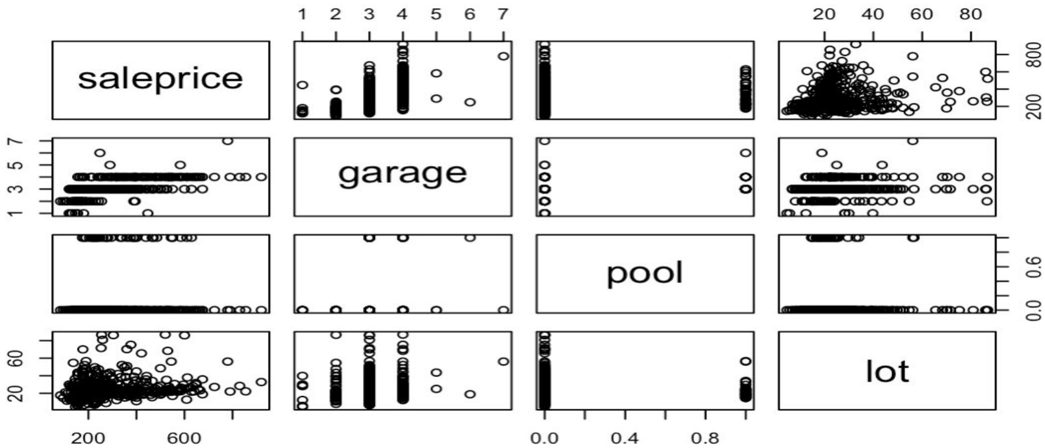
In order to determine whether variables relating to the size of the home affect the sales prices of a home, we will first be sure that we factor numerical variables, such as garage and pool, to be categorical predictors, then compare the scatterplots of the relevant variables against sales price. We will then look at the p-values of each variable and check for their significance under our chosen $\alpha = 0.05$ level. Afterwards, we will test our chosen predictors using stepwise regression as well as best subsets regression procedures to determine the overall best model. Once we have our final linear model, we will test its linearity by looking at the normal Q-Q plot and Residual versus Fit plot. Finally, using the best linear model, we will conduct a hypothesis test.

Regression Analysis, Results and Interpretation

Since we are comparing the effects of square feet, beds, baths, garage, pool, and lot against the sales price, this allowed us to visually see the relationship each individual predictor has with sales price, as seen in Figure 1 below. We saw that there is a relationship between all of the listed variables and sales price based on the scatterplots.

Figure 1





We confirmed these relationships with the numerical values yielded by the summary of a linear model, which is shown in Figure 2 below. From the summary table, we see that all variables except garage and pool are significant under our $\alpha = 0.05$ level. Looking at the multiple garage variables, we see that at least some of them are significant, and can therefore include the predictor in its entirety. Shifting our focus to the pool variable, we see that it is not significant at our $\alpha = 0.05$, and can therefore be excluded from future tests.

Figure 2

```
    Min      1Q  Median      3Q     Max
-194.55  -32.21   -3.86   24.70  343.99

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -26.6532    28.5085  -0.935 0.350272
sqfeet      113.4385     7.4513  15.224  < 2e-16 ***
beds         -9.3287     3.9565  -2.358 0.018761 *
baths        22.3518     4.7969   4.660 4.05e-06 ***
garage1     -25.4294    28.5891  -0.889 0.374166
garage2     -18.5329    27.4570  -0.675 0.499997
garage3      59.9729    28.8230   2.081 0.037958 *
garage4      -5.2934    57.5419  -0.092 0.926740
garage5      -7.0501    76.7064  -0.092 0.926806
garage7     178.3557    77.3771   2.305 0.021567 *
pool1        11.2388    12.6229   0.890 0.373701
lot           0.9366     0.2712   3.453 0.000601 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 70.49 on 509 degrees of freedom
Multiple R-squared:  0.7432,	Adjusted R-squared:  0.7376
F-statistic: 133.9 on 11 and 509 DF,  p-value: < 2.2e-16
```

From the stepwise regression procedure in Figure 3 and 4 below, we see that the model containing the predictors square feet, garage, baths, and lot, in that specific order, yields the smallest AIC value, and is therefore the “best” model.

Figure 3

```
Step:  AIC=4444.88
saleprice ~ sqfeet + garage + baths + lot + beds

         Df Sum of Sq     RSS    AIC
<none>                2533115 4444.9
+ pool    1      3939 2529177 4446.1
- beds    1     27273 2560389 4448.5
- lot     1     57592 2590708 4454.6
- baths   1    111674 2644790 4465.4
- garage  6    388143 2921258 4507.2
- sqfeet  1   1163261 3696376 4639.8

Call:
lm(formula = saleprice ~ sqfeet + garage + baths + lot + beds)

Coefficients:
(Intercept)       sqfeet      garage1      garage2      garage3      garage4
   -27.5121     113.8199     -25.5754     -18.3717      59.7833      -6.6284
    garage5      garage7        baths          lot         beds
     3.3654     176.5033      22.6755       0.9217      -9.2680
```

Figure 4

```
Call:
lm(formula = saleprice ~ sqfeet + garage + baths + lot + beds)

Coefficients:
(Intercept)       sqfeet      garage1      garage2      garage3      garage4
   -27.5121     113.8199     -25.5754     -18.3717      59.7833      -6.6284
    garage5      garage7        baths          lot         beds
     3.3654     176.5033      22.6755       0.9217      -9.2680


Call:
lm(formula = saleprice ~ sqfeet + garage + baths + lot + beds)

Residuals:
    Min      1Q  Median      3Q     Max
-185.49  -33.22   -3.20   24.98  342.19

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -27.5121    28.4864  -0.966 0.334603
sqfeet      113.8199     7.4374  15.304  < 2e-16 ***
garage1     -25.5754    28.5828  -0.895 0.371326
garage2     -18.3717    27.4509  -0.669 0.503633
garage3      59.7833    28.8164   2.075 0.038522 *
garage4      -6.6284    57.5107  -0.115 0.908288
garage5       3.3654    75.7938   0.044 0.964602
garage7     176.5033    77.3334   2.282 0.022878 *
baths        22.6755     4.7821   4.742 2.75e-06 ***
lot           0.9217     0.2707   3.405 0.000713 ***
beds         -9.2680     3.9551  -2.343 0.019497 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 70.48 on 510 degrees of freedom
Multiple R-squared:  0.7428,	Adjusted R-squared:  0.7377
F-statistic: 147.3 on 10 and 510 DF,  p-value: < 2.2e-16
```

From the first row of our Best Subsets Regression procedure in Figure 5 below, we see that the lowest mean square error (MSE) occurs when we have 5 predictors in the model. The adjusted R squared, found in the third row, is the largest when there are 5 predictors in the model. When comparing the models from both the Stepwise Regression and Best Subsets Regression procedures, it is apparent that both result in the same "best" model. Therefore, our final model includes the variables square feet, garage, baths, lot, and beds.

Figure 5

```
[1] 6129.895 5656.511 5521.704 5417.445 5316.630
[1] 0.6769435 0.7024661 0.7101177 0.7161413 0.7219636 0.7220565
[1] 0.6763211 0.7013173 0.7084356 0.7139408 0.7192642 0.7188120
  (Intercept) sqfeet  beds baths garage  pool   lot
1        TRUE   TRUE FALSE FALSE  FALSE FALSE FALSE
2        TRUE   TRUE FALSE FALSE   TRUE FALSE FALSE
3        TRUE   TRUE FALSE  TRUE   TRUE FALSE FALSE
4        TRUE   TRUE FALSE  TRUE   TRUE FALSE  TRUE
5        TRUE   TRUE  TRUE  TRUE   TRUE FALSE  TRUE
6        TRUE   TRUE  TRUE  TRUE   TRUE  TRUE  TRUE
```

Initially, our R squared value was 0.7377, and we tested to see if performing a log transformation on each individual variable would increase the coefficient of determination. This resulted in a final model using the log transformations of sales price, square feet, and baths. Our new coefficient of determination is 0.7884.

Now that we have obtained our final model, we confirmed that all four of the "**LINE**" conditions have been met. Our model satisfies the **L**inear condition because the residuals "bounce randomly" around the zero line in Figure 6 below. Our model also satisfies the **E**qual variance condition because it forms a rough "horizontal band" around the zero line. Since we do not know the order in which the data was collected, we cannot use the Residual versus Order plot to check the **I**ndependence condition. To check the **N**ormal condition, we looked at the Q-Q plot in Figure 6 below. Since the results are predominantly linear, we can assume that the error is normally distributed. Thus, our final linear model satisfies the "**LINE**" criterion.

Figure 6

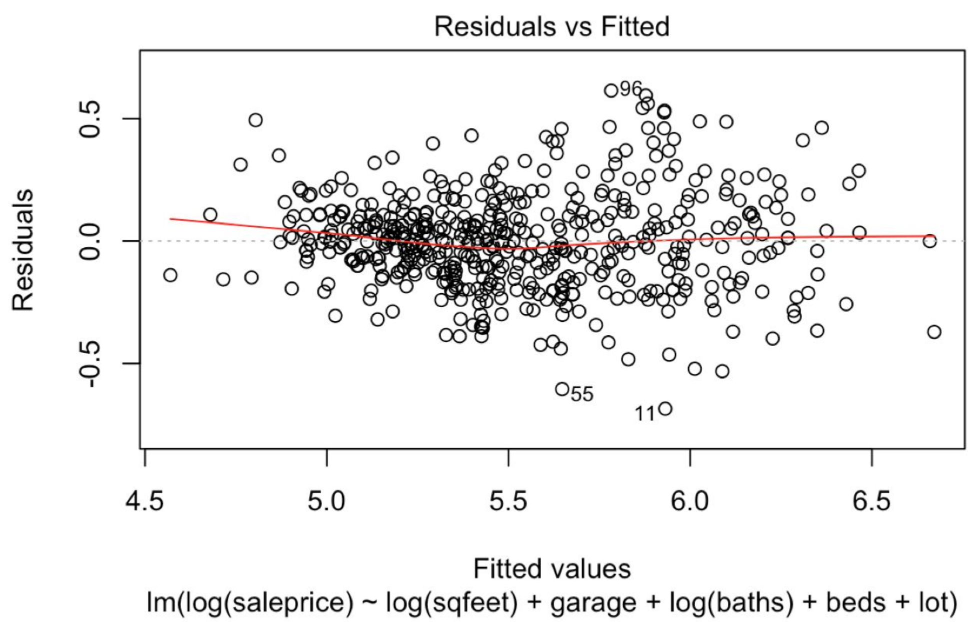
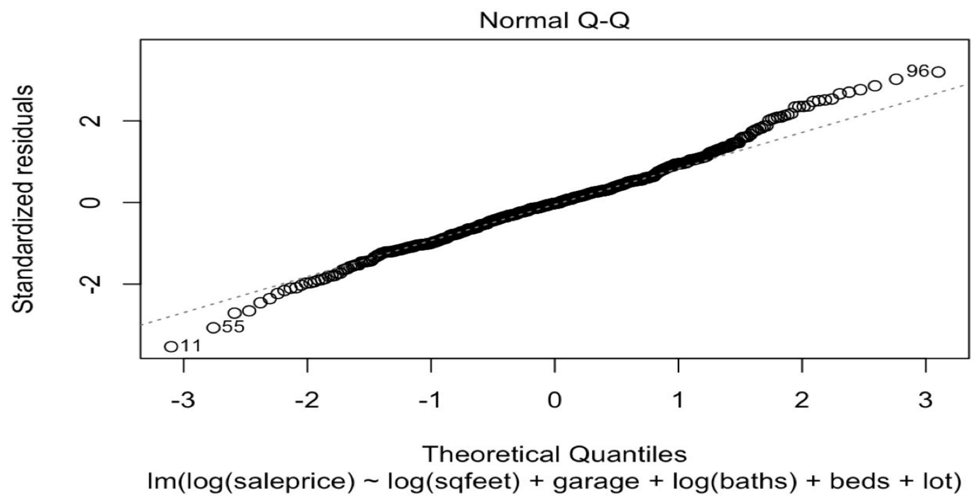


Figure 7



Lastly, we conducted a hypothesis test for our final model. In this test, H_0 : β1 = β2 = β3 = β4 = β5 = 0 and H_1 : At least one βk ≠ 0 (k = 1, 2, 3, 4, 5) where β1 = slope of log(square feet), β2 = slope of garage, β3 = slope of log(baths), β4 = slope of beds, and β5 = slope of lots.

Looking at Figure 8 below, we find that garage1, garage2, and beds have a negative relationship with sales price, but since they are insignificant compared to our other predictors, the overall relationship between sales price and these predictors remains positive. Lastly, looking at our final p-value, we find that it is less than our $\alpha = 0.05$ level of significance, and therefore fail to accept the null hypothesis.

Figure 8

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.68427 -0.12670 -0.00457  0.10649  0.61440

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.6402998  0.0788307  58.864  < 2e-16 ***
log(sqfeet)  0.8570288  0.0499013  17.174  < 2e-16 ***
garage1     -0.0985848  0.0803507  -1.227   0.2204
garage2     -0.0400975  0.0773739  -0.518   0.6045
garage3      0.1677671  0.0814413   2.060   0.0399 *
garage4      0.0045986  0.1617527   0.028   0.9773
garage5      0.0433414  0.2132169   0.203   0.8390
garage7      0.2785320  0.2165690   1.286   0.1990
log(baths)   0.2047532  0.0310581   6.593 1.08e-10 ***
beds        -0.0078404  0.0111067  -0.706   0.4806
lot          0.0030077  0.0007622   3.946 9.05e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1982 on 510 degrees of freedom
Multiple R-squared:  0.7925,    Adjusted R-squared:  0.7884
F-statistic: 194.7 on 10 and 510 DF,  p-value: < 2.2e-16
```

Conclusion

In conclusion, the variables pertaining to size, including the finished area of residence, the number of cars that the garage will hold, total number of bathrooms in residence, number of bedrooms in residence, and lot size do affect the sales price of a home. Overall, we see that the sum of the variables relating to size increase the sales price of a home. However, looking at the slope of garage1, garage2, and beds, we see that not all of the variables in the model are positively correlated, which tells us that large variable values do not always result in high sales prices. To improve our analysis, we could include predictors not in the data set, such as number of residents, number of floors, and rooms with unspecified purposes, such as basements or studies.

Appendix

1. Calling/Declaring library, data set, and variables
```{r}
realestate=read.table("realestate.txt", header=TRUE)
saleprice=realestate$SalePrice
sqfeet=realestate$SqFeet
beds=realestate$Beds
baths=realestate$Baths
garage=factor(realestate$Garage)
pool=factor(realestate$Pool)
lot=realestate$Lot
```
2. Plotting all relevant variables
```{r}
pairs(~saleprice+sqfeet+beds+baths)
pairs(~saleprice+garage+pool+lot)
fitfirst=lm(saleprice~sqfeet+beds+baths+garage+pool+lot)
summary(fitfirst)S
```
3. Stepwise regression test
```{r}
mod0=lm(saleprice~1)
mod.upper=lm(saleprice~sqfeet+beds+baths+garage+pool+lot)
step(mod0,scope=list(lower=mod0, upper=mod.upper))
mod.final=lm(saleprice~sqfeet+garage+baths+lot+beds)
summary(mod.final)
```
4. Subset regression test
```{r}
library(leaps)
```

````
mod=regsubsets(cbind(sqfeet,beds,baths,garage,pool,lot),saleprice)
summary.mod=summary(mod)
n=521
rss=summary.mod$rss
mses=c(rss[1]/(n-2),rss[2]/(n-3),rss[3]/(n-4),rss[4]/(n-5),rss[5]/(n-6))
mses
summary.mod$rsq
summary.mod$adjr2
summary.mod$which
```
````

5. Final model: Checking for linearity and hypothesis testing

````
```{r}
final=lm(log(saleprice)~log(sqfeet)+garage+log(baths)+beds+lot)
plot(final)
summary(final)
```
````
````