NY Housing

**1.a** I plan on looking at how price is determined using a multivariate regression taking real estate data into account. The features I want to take into account are gross square feet, the neighborhood, and building class. I also want to look at the relationship of sale-date and price across the entire dataset to identify patterns in the macroeconomic housing market. A logarithmic scale makes the price more manageable.

## Exploration

In the exploration, I looked at the distributions of price and square-foot variables and tentavely tested for a relationship. I decided to use a log-scale for price. I also looked at the log of sale price over time.

summary(bronx$price)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 10500 200000 350000 827600 480000 137300000

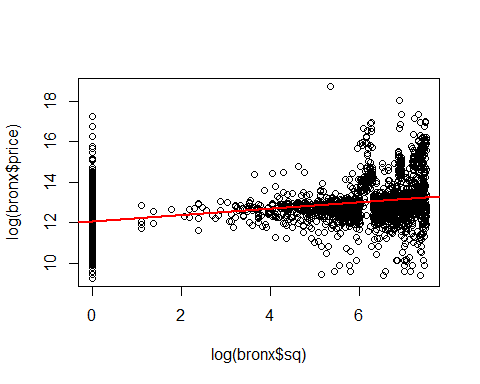
summary(bronx$sq)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.0 1.0 344.0 528.5 931.0 1817.0

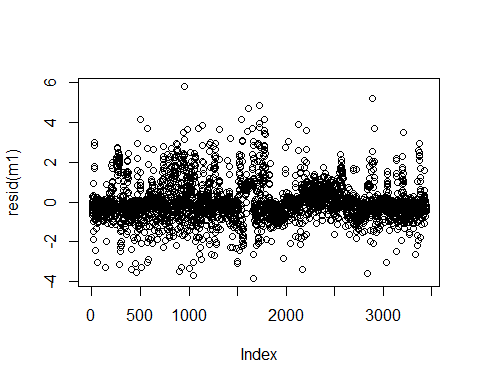
plot(log(bronx$sq), log(bronx$price))  
m1<-lm(log(bronx$price)~log(bronx$sq))  
summary(m1)

##   
## Call:  
## lm(formula = log(bronx$price) ~ log(bronx$sq))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.8418 -0.4810 -0.1194 0.2991 5.8175   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 12.089249 0.028219 428.41 <2e-16 \*\*\*  
## log(bronx$sq) 0.155129 0.005355 28.97 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9629 on 3443 degrees of freedom  
## Multiple R-squared: 0.196, Adjusted R-squared: 0.1957   
## F-statistic: 839.1 on 1 and 3443 DF, p-value: < 2.2e-16

plot(log(bronx$sq), log(bronx$price))  
abline(m1,col="red",lwd=2)



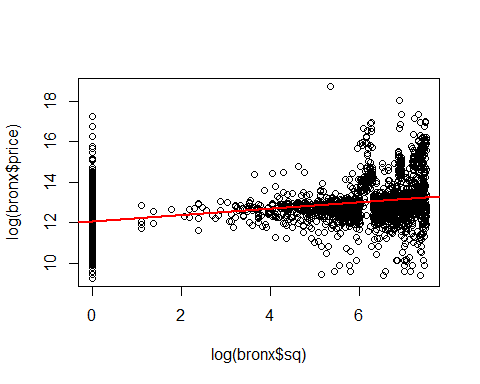
plot(resid(m1))



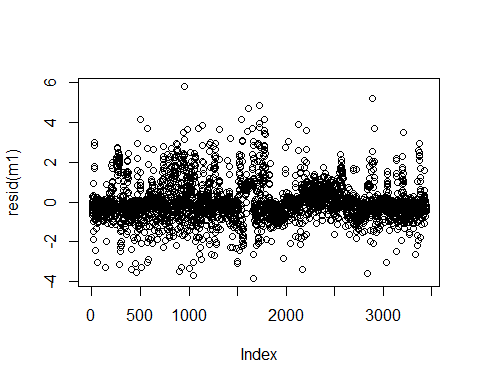
bronx$NEIGHBORHOOD <- as.numeric(bronx$NEIGHBORHOOD)  
bronx$BUILDING.CLASS.CATEGORY <- as.numeric(bronx$BUILDING.CLASS.CATEGORY)  
summary(m1)

##   
## Call:  
## lm(formula = log(bronx$price) ~ log(bronx$sq))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.8418 -0.4810 -0.1194 0.2991 5.8175   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 12.089249 0.028219 428.41 <2e-16 \*\*\*  
## log(bronx$sq) 0.155129 0.005355 28.97 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9629 on 3443 degrees of freedom  
## Multiple R-squared: 0.196, Adjusted R-squared: 0.1957   
## F-statistic: 839.1 on 1 and 3443 DF, p-value: < 2.2e-16

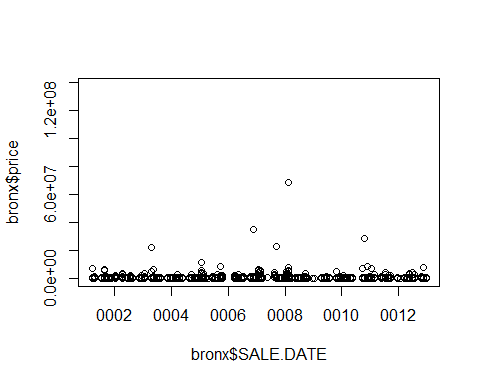
plot(log(bronx$sq), log(bronx$price))  
abline(m1,col="red",lwd=2)



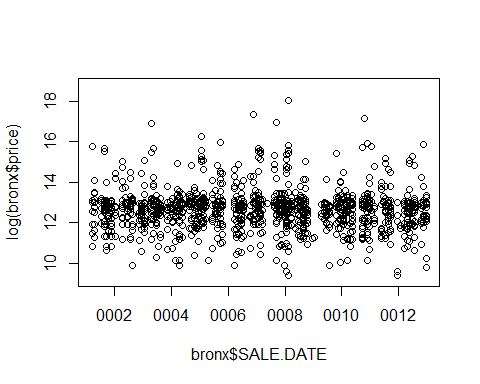
plot(resid(m1))



plot(bronx$SALE.DATE, bronx$price)



plot(bronx$SALE.DATE, log(bronx$price))

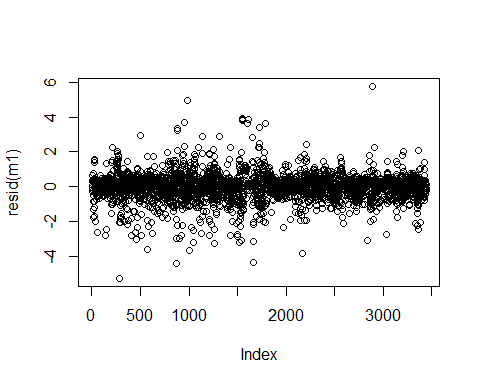


**1.b.**

It is necessary to clean the data. Currency needs to be converted to a numeric type, as does square feet. It is necessary to filter out 0 price and 0 square feet observations. Neighborhood and building category need to be treated as factors.

convertCurrency <- function(currency) {  
 currency1 <- sub('$','',as.character(currency),fixed=TRUE)  
 currency2 <- as.numeric(gsub('\\,','',as.character(currency1)))   
 currency2  
}  
bronx <- unique(read.csv("~/data/data/rollingsales\_bronx.csv"))  
bronx <- bronx[which(bronx$sq>0 & bronx$sq>0 & bronx$price>10000),]  
bronx$price <- convertCurrency(bronx$SALE.PRICE)  
bronx$sq <- as.numeric(bronx$GROSS.SQUARE.FEET)  
bronx$lsq <-as.numeric(gsub(",","", bronx$LAND.SQUARE.FEET))  
bronx$SALE.DATE <- as.Date(gsub("[^]:digit:]]","",bronx$SALE.DATE))  
bronx$YEAR.BUILT <- as.numeric(gsub("[^]:digit:]]","",bronx$YEAR.BUILT))  
bronx$NEIGHBORHOOD <- as.factor(bronx$NEIGHBORHOOD)

m1<-lm(log(bronx$price)~bronx$sq+factor(bronx$BUILDING.CLASS.CATEGORY)+factor(bronx$NEIGHBORHOOD))  
plot(resid(m1))



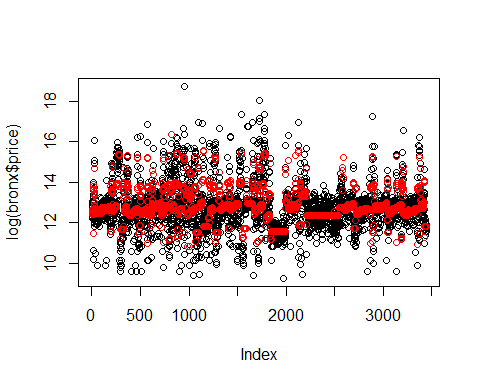
summary(m1)

##   
## Call:  
## lm(formula = log(bronx$price) ~ bronx$sq + factor(bronx$BUILDING.CLASS.CATEGORY) +   
## factor(bronx$NEIGHBORHOOD))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.2647 -0.3098 0.0685 0.3297 5.7610   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) 1.103e+01 3.353e-01 32.895  
## bronx$sq 1.105e-04 3.685e-05 2.997  
## factor(bronx$BUILDING.CLASS.CATEGORY)2 1.363e+00 3.070e-01 4.440  
## factor(bronx$BUILDING.CLASS.CATEGORY)3 1.479e+00 3.073e-01 4.813  
## factor(bronx$BUILDING.CLASS.CATEGORY)4 1.527e+00 3.107e-01 4.916  
## factor(bronx$BUILDING.CLASS.CATEGORY)5 1.589e+00 3.260e-01 4.874  
## factor(bronx$BUILDING.CLASS.CATEGORY)6 1.376e+00 3.248e-01 4.236  
## factor(bronx$BUILDING.CLASS.CATEGORY)7 -6.895e-02 4.038e-01 -0.171  
## factor(bronx$BUILDING.CLASS.CATEGORY)8 2.576e+00 3.116e-01 8.269  
## factor(bronx$BUILDING.CLASS.CATEGORY)9 4.015e+00 3.199e-01 12.550  
## factor(bronx$BUILDING.CLASS.CATEGORY)10 2.595e-02 3.241e-01 0.080  
## factor(bronx$BUILDING.CLASS.CATEGORY)11 4.649e-01 3.048e-01 1.525  
## factor(bronx$BUILDING.CLASS.CATEGORY)12 2.628e+00 5.776e-01 4.550  
## factor(bronx$BUILDING.CLASS.CATEGORY)13 8.397e-01 3.836e-01 2.189  
## factor(bronx$BUILDING.CLASS.CATEGORY)14 8.634e-01 3.127e-01 2.762  
## factor(bronx$BUILDING.CLASS.CATEGORY)15 1.901e+00 3.363e-01 5.653  
## factor(bronx$BUILDING.CLASS.CATEGORY)17 1.673e+00 3.208e-01 5.216  
## factor(bronx$BUILDING.CLASS.CATEGORY)18 2.311e+00 3.623e-01 6.379  
## factor(bronx$BUILDING.CLASS.CATEGORY)19 2.664e+00 3.210e-01 8.297  
## factor(bronx$BUILDING.CLASS.CATEGORY)20 2.422e+00 3.748e-01 6.463  
## factor(bronx$BUILDING.CLASS.CATEGORY)21 6.272e-01 4.797e-01 1.307  
## factor(bronx$BUILDING.CLASS.CATEGORY)22 2.305e+00 3.175e-01 7.260  
## factor(bronx$BUILDING.CLASS.CATEGORY)23 3.509e+00 3.323e-01 10.561  
## factor(bronx$BUILDING.CLASS.CATEGORY)24 2.494e+00 3.469e-01 7.191  
## factor(bronx$BUILDING.CLASS.CATEGORY)25 3.264e+00 4.825e-01 6.765  
## factor(bronx$BUILDING.CLASS.CATEGORY)26 3.866e+00 6.095e-01 6.343  
## factor(bronx$BUILDING.CLASS.CATEGORY)27 1.862e+00 6.088e-01 3.058  
## factor(bronx$BUILDING.CLASS.CATEGORY)28 2.642e+00 8.124e-01 3.252  
## factor(bronx$BUILDING.CLASS.CATEGORY)29 1.966e+00 4.179e-01 4.705  
## factor(bronx$BUILDING.CLASS.CATEGORY)30 2.561e+00 8.089e-01 3.166  
## factor(bronx$BUILDING.CLASS.CATEGORY)31 2.508e+00 8.058e-01 3.112  
## factor(bronx$BUILDING.CLASS.CATEGORY)32 1.792e+00 3.765e-01 4.759  
## factor(bronx$NEIGHBORHOOD)2 9.083e-02 1.507e-01 0.603  
## factor(bronx$NEIGHBORHOOD)3 1.560e-01 1.568e-01 0.995  
## factor(bronx$NEIGHBORHOOD)4 9.317e-02 1.803e-01 0.517  
## factor(bronx$NEIGHBORHOOD)5 1.014e+00 7.690e-01 1.318  
## factor(bronx$NEIGHBORHOOD)6 3.976e-01 5.431e-01 0.732  
## factor(bronx$NEIGHBORHOOD)7 -1.767e-02 1.553e-01 -0.114  
## factor(bronx$NEIGHBORHOOD)8 5.914e-03 1.652e-01 0.036  
## factor(bronx$NEIGHBORHOOD)9 1.236e+00 4.509e-01 2.742  
## factor(bronx$NEIGHBORHOOD)10 -2.127e-02 1.804e-01 -0.118  
## factor(bronx$NEIGHBORHOOD)11 5.480e-01 3.706e-01 1.479  
## factor(bronx$NEIGHBORHOOD)12 2.970e-01 1.632e-01 1.820  
## factor(bronx$NEIGHBORHOOD)13 -4.209e-02 1.969e-01 -0.214  
## factor(bronx$NEIGHBORHOOD)14 1.314e-01 1.691e-01 0.777  
## factor(bronx$NEIGHBORHOOD)15 1.163e+00 2.255e-01 5.157  
## factor(bronx$NEIGHBORHOOD)16 1.971e-01 1.701e-01 1.158  
## factor(bronx$NEIGHBORHOOD)17 6.376e-03 1.606e-01 0.040  
## factor(bronx$NEIGHBORHOOD)18 -2.059e-01 1.747e-01 -1.179  
## factor(bronx$NEIGHBORHOOD)19 1.708e-02 1.675e-01 0.102  
## factor(bronx$NEIGHBORHOOD)20 3.508e-01 1.581e-01 2.219  
## factor(bronx$NEIGHBORHOOD)21 3.810e-02 1.595e-01 0.239  
## factor(bronx$NEIGHBORHOOD)22 2.290e-01 1.581e-01 1.449  
## factor(bronx$NEIGHBORHOOD)23 2.376e-01 1.500e-01 1.584  
## factor(bronx$NEIGHBORHOOD)24 1.328e-02 1.684e-01 0.079  
## factor(bronx$NEIGHBORHOOD)25 1.761e-01 1.727e-01 1.020  
## factor(bronx$NEIGHBORHOOD)26 -2.964e-01 1.600e-01 -1.852  
## factor(bronx$NEIGHBORHOOD)27 5.909e-01 2.033e-01 2.906  
## factor(bronx$NEIGHBORHOOD)28 2.610e-01 1.650e-01 1.582  
## factor(bronx$NEIGHBORHOOD)29 3.586e-01 1.762e-01 2.035  
## factor(bronx$NEIGHBORHOOD)30 8.872e-01 1.489e-01 5.958  
## factor(bronx$NEIGHBORHOOD)31 2.394e-01 1.549e-01 1.546  
## factor(bronx$NEIGHBORHOOD)32 1.725e-02 1.509e-01 0.114  
## factor(bronx$NEIGHBORHOOD)33 2.662e-01 1.551e-01 1.716  
## factor(bronx$NEIGHBORHOOD)34 5.012e-02 3.354e-01 0.149  
## factor(bronx$NEIGHBORHOOD)35 8.241e-02 1.612e-01 0.511  
## factor(bronx$NEIGHBORHOOD)36 1.971e-01 1.672e-01 1.179  
## factor(bronx$NEIGHBORHOOD)37 -1.336e-02 1.511e-01 -0.088  
## factor(bronx$NEIGHBORHOOD)38 3.225e-01 1.739e-01 1.855  
## Pr(>|t|)   
## (Intercept) < 2e-16 \*\*\*  
## bronx$sq 0.00274 \*\*   
## factor(bronx$BUILDING.CLASS.CATEGORY)2 9.27e-06 \*\*\*  
## factor(bronx$BUILDING.CLASS.CATEGORY)3 1.55e-06 \*\*\*  
## factor(bronx$BUILDING.CLASS.CATEGORY)4 9.27e-07 \*\*\*  
## factor(bronx$BUILDING.CLASS.CATEGORY)5 1.14e-06 \*\*\*  
## factor(bronx$BUILDING.CLASS.CATEGORY)6 2.34e-05 \*\*\*  
## factor(bronx$BUILDING.CLASS.CATEGORY)7 0.86442   
## factor(bronx$BUILDING.CLASS.CATEGORY)8 < 2e-16 \*\*\*  
## factor(bronx$BUILDING.CLASS.CATEGORY)9 < 2e-16 \*\*\*  
## factor(bronx$BUILDING.CLASS.CATEGORY)10 0.93619   
## factor(bronx$BUILDING.CLASS.CATEGORY)11 0.12733   
## factor(bronx$BUILDING.CLASS.CATEGORY)12 5.54e-06 \*\*\*  
## factor(bronx$BUILDING.CLASS.CATEGORY)13 0.02868 \*   
## factor(bronx$BUILDING.CLASS.CATEGORY)14 0.00578 \*\*   
## factor(bronx$BUILDING.CLASS.CATEGORY)15 1.71e-08 \*\*\*  
## factor(bronx$BUILDING.CLASS.CATEGORY)17 1.94e-07 \*\*\*  
## factor(bronx$BUILDING.CLASS.CATEGORY)18 2.03e-10 \*\*\*  
## factor(bronx$BUILDING.CLASS.CATEGORY)19 < 2e-16 \*\*\*  
## factor(bronx$BUILDING.CLASS.CATEGORY)20 1.17e-10 \*\*\*  
## factor(bronx$BUILDING.CLASS.CATEGORY)21 0.19116   
## factor(bronx$BUILDING.CLASS.CATEGORY)22 4.76e-13 \*\*\*  
## factor(bronx$BUILDING.CLASS.CATEGORY)23 < 2e-16 \*\*\*  
## factor(bronx$BUILDING.CLASS.CATEGORY)24 7.90e-13 \*\*\*  
## factor(bronx$BUILDING.CLASS.CATEGORY)25 1.56e-11 \*\*\*  
## factor(bronx$BUILDING.CLASS.CATEGORY)26 2.55e-10 \*\*\*  
## factor(bronx$BUILDING.CLASS.CATEGORY)27 0.00225 \*\*   
## factor(bronx$BUILDING.CLASS.CATEGORY)28 0.00116 \*\*   
## factor(bronx$BUILDING.CLASS.CATEGORY)29 2.64e-06 \*\*\*  
## factor(bronx$BUILDING.CLASS.CATEGORY)30 0.00156 \*\*   
## factor(bronx$BUILDING.CLASS.CATEGORY)31 0.00187 \*\*   
## factor(bronx$BUILDING.CLASS.CATEGORY)32 2.03e-06 \*\*\*  
## factor(bronx$NEIGHBORHOOD)2 0.54664   
## factor(bronx$NEIGHBORHOOD)3 0.32001   
## factor(bronx$NEIGHBORHOOD)4 0.60543   
## factor(bronx$NEIGHBORHOOD)5 0.18758   
## factor(bronx$NEIGHBORHOOD)6 0.46420   
## factor(bronx$NEIGHBORHOOD)7 0.90938   
## factor(bronx$NEIGHBORHOOD)8 0.97145   
## factor(bronx$NEIGHBORHOOD)9 0.00614 \*\*   
## factor(bronx$NEIGHBORHOOD)10 0.90614   
## factor(bronx$NEIGHBORHOOD)11 0.13928   
## factor(bronx$NEIGHBORHOOD)12 0.06892 .   
## factor(bronx$NEIGHBORHOOD)13 0.83071   
## factor(bronx$NEIGHBORHOOD)14 0.43726   
## factor(bronx$NEIGHBORHOOD)15 2.65e-07 \*\*\*  
## factor(bronx$NEIGHBORHOOD)16 0.24677   
## factor(bronx$NEIGHBORHOOD)17 0.96833   
## factor(bronx$NEIGHBORHOOD)18 0.23852   
## factor(bronx$NEIGHBORHOOD)19 0.91877   
## factor(bronx$NEIGHBORHOOD)20 0.02655 \*   
## factor(bronx$NEIGHBORHOOD)21 0.81118   
## factor(bronx$NEIGHBORHOOD)22 0.14742   
## factor(bronx$NEIGHBORHOOD)23 0.11332   
## factor(bronx$NEIGHBORHOOD)24 0.93715   
## factor(bronx$NEIGHBORHOOD)25 0.30779   
## factor(bronx$NEIGHBORHOOD)26 0.06408 .   
## factor(bronx$NEIGHBORHOOD)27 0.00368 \*\*   
## factor(bronx$NEIGHBORHOOD)28 0.11382   
## factor(bronx$NEIGHBORHOOD)29 0.04194 \*   
## factor(bronx$NEIGHBORHOOD)30 2.82e-09 \*\*\*  
## factor(bronx$NEIGHBORHOOD)31 0.12219   
## factor(bronx$NEIGHBORHOOD)32 0.90898   
## factor(bronx$NEIGHBORHOOD)33 0.08619 .   
## factor(bronx$NEIGHBORHOOD)34 0.88119   
## factor(bronx$NEIGHBORHOOD)35 0.60924   
## factor(bronx$NEIGHBORHOOD)36 0.23845   
## factor(bronx$NEIGHBORHOOD)37 0.92954   
## factor(bronx$NEIGHBORHOOD)38 0.06374 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7401 on 3376 degrees of freedom  
## Multiple R-squared: 0.5343, Adjusted R-squared: 0.5249   
## F-statistic: 56.97 on 68 and 3376 DF, p-value: < 2.2e-16

Our adjusted R-squared is 0.5249, meaning this model explains more than half of the variance in sale price. According to the p-value (p-value < 2.2e-16) this model is highly significant. Many of the building-class coefficients are not significant, but some are highly significant.

**2. a**

## Predicted prices



The predicted prices look pretty good. The model captured the pattern around the 2000 index very well. A local model would be better at explaining the outliers.

**2. b** The adjusted R-squared is good, and the model is highly significant (p-value < 2.2e-16). Not all the building classifications are significant.

Residual standard error: 0.7401 on 3376 degrees of freedom  
Multiple R-squared: 0.5343, Adjusted R-squared: 0.5249   
F-statistic: 56.97 on 68 and 3376 DF, p-value: < 2.2e-16

You can see that the model fits the data decently from the adjusted R-squared. The F-statistic is an appropriate test statistic for this model, and it shows that model is significant.

**2. c** There is a lot of data in the dataset we don't want in our model. Those include records that do not represent a sale on the open market. There should be some explanation for the 0-square-feet and no-sale-price properties. It would take research and domain knowledge to go further into it.

**3** The model used was appropriate for learning the global patterns, but a local model would better understand the outliers near the bottom and the top of the price range. A non-linear transform on the square-foot variable might make sense intuitively. Due to the neighborhood and category factors, the model is high-dimensional. More data would make it better and give us some more significant estimators on the some of the coefficients.