Statistics Mini Project with Housing Data

Raghav Joshi

April 11, 2015

The following write-up documents the steps taken and the assumptions made to create the final model.

The first step was to set the working directory and read the csv file into R. This can be seen below. Just to take extra precautions, any preexisting object references were removed from the global environment

rm(list=ls())  
data <- read.csv("Housing Price.csv")  
data <- na.omit(data)

Before delving into any analysis, a quick summary was done to see which covariates were present and whether there were any indicators as to whether there were any covariates with strong multicollinearity. This information could be useful when making certain assumptions about the final model.

## 'data.frame': 6777 obs. of 9 variables:  
## $ PROPERTY\_ID: int 101825925 101879091 101879098 101879104 101879130 104829818 104829819 109100605 109100606 109100607 ...  
## $ ZIP : int 90621 90621 90621 90621 90621 90621 90621 90621 90621 90621 ...  
## $ PRICE : int 450000 179500 196500 236000 279500 275000 550000 635000 630000 622000 ...  
## $ year : int 2004 2007 2007 2009 2006 2008 2005 2005 2008 2008 ...  
## $ quarter : int 1 4 4 1 2 2 1 2 2 2 ...  
## $ bathrooms : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ bedrooms : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ YR\_BLT : Factor w/ 85 levels ".","1894","1898",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ SQFT : int 0 0 0 0 0 0 0 0 0 0 ...

## PROPERTY\_ID ZIP PRICE year   
## Min. : 38623263 Min. :90621 Min. : 6000 Min. :1988   
## 1st Qu.: 38625781 1st Qu.:90621 1st Qu.: 190000 1st Qu.:1996   
## Median : 38840972 Median :90621 Median : 273000 Median :2000   
## Mean : 39332330 Mean :90622 Mean : 328193 Mean :2000   
## 3rd Qu.: 38849331 3rd Qu.:90623 3rd Qu.: 425000 3rd Qu.:2004   
## Max. :124187083 Max. :90623 Max. :2058000 Max. :2009   
##   
## quarter bathrooms bedrooms YR\_BLT   
## Min. :1.000 Min. : 0.000 Min. :0.000 1966 : 701   
## 1st Qu.:2.000 1st Qu.: 1.000 1st Qu.:2.000 1973 : 564   
## Median :2.000 Median : 3.000 Median :3.000 1965 : 534   
## Mean :2.511 Mean : 2.833 Mean :3.018 1969 : 375   
## 3rd Qu.:3.000 3rd Qu.: 4.000 3rd Qu.:4.000 1971 : 370   
## Max. :4.000 Max. :12.000 Max. :9.000 . : 300   
## (Other):3933   
## SQFT   
## Min. : 0   
## 1st Qu.:1126   
## Median :1565   
## Mean :1597   
## 3rd Qu.:2114   
## Max. :8956   
##

## Descriptive Statistics and Initial Observations

First it's apparent that there are 6777 observations with 11 covariates. The main observations are as follows:

### 1)

PROPERTY\_ID seems extraneous to any analysis so we can likely proceed without it.

### 2)

There seem to be only two zip-codes so perhaps a conversion into a categorical variable could be useful later

### 3)

The average price is greater than the median price so there is probably a positive skew and thus a log transformation might be appropriate (this can later be confirmed when we construct a histogram of the price).

### 4)

Out of the 6777 observations, 300 have a "." associated with them for the YR\_BLT covariate. In addition, it appears that when a "." is noted with this covariate, the number of bedrooms, bathrooms, and the square feet is 0 thus indicating that this property has only land and no house. This could be a problem when predicting price.

new\_data <- read.csv("new\_data.csv")  
str(new\_data)

## 'data.frame': 6477 obs. of 9 variables:  
## $ PROPERTY\_ID: int 38627962 38849614 38849225 38849634 38849634 38623545 38623545 38838011 38849636 38849823 ...  
## $ ZIP : int 90621 90621 90621 90621 90621 90621 90621 90623 90621 90621 ...  
## $ PRICE : int 185000 105000 95000 155000 98500 510000 260000 1350000 96000 122000 ...  
## $ year : int 2001 1997 1997 1990 1997 2006 2009 2005 1997 1991 ...  
## $ quarter : int 3 4 4 4 2 2 2 1 3 1 ...  
## $ bathrooms : int 0 1 1 1 1 1 1 2 1 1 ...  
## $ bedrooms : int 0 2 2 2 2 3 3 4 2 2 ...  
## $ YR\_BLT : int 1894 1898 1902 1902 1902 1903 1903 1903 1904 1904 ...  
## $ SQFT : int 952 1308 930 1046 1046 1260 1260 2741 912 636 ...

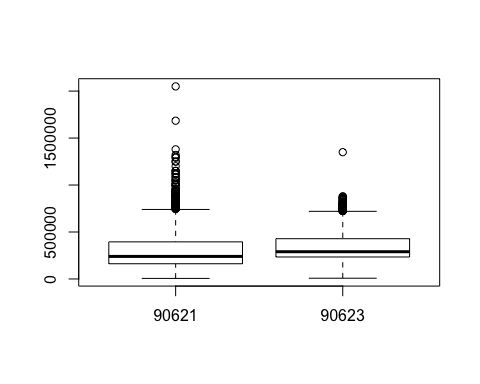
summary(new\_data)

## PROPERTY\_ID ZIP PRICE year   
## Min. : 38623263 Min. :90621 Min. : 6000 Min. :1988   
## 1st Qu.: 38625777 1st Qu.:90621 1st Qu.: 188000 1st Qu.:1996   
## Median : 38840183 Median :90621 Median : 267000 Median :2000   
## Mean : 38840256 Mean :90622 Mean : 318302 Mean :2000   
## 3rd Qu.: 38849158 3rd Qu.:90623 3rd Qu.: 409000 3rd Qu.:2004   
## Max. :109189813 Max. :90623 Max. :2050000 Max. :2009   
## quarter bathrooms bedrooms YR\_BLT   
## Min. :1.000 Min. :0.000 Min. :0.000 Min. :1894   
## 1st Qu.:2.000 1st Qu.:2.000 1st Qu.:3.000 1st Qu.:1957   
## Median :3.000 Median :3.000 Median :3.000 Median :1966   
## Mean :2.516 Mean :2.962 Mean :3.156 Mean :1963   
## 3rd Qu.:3.000 3rd Qu.:4.000 3rd Qu.:4.000 3rd Qu.:1972   
## Max. :4.000 Max. :9.000 Max. :7.000 Max. :2005   
## SQFT   
## Min. : 0   
## 1st Qu.:1240   
## Median :1603   
## Mean :1671   
## 3rd Qu.:2118   
## Max. :8956

attach(new\_data)

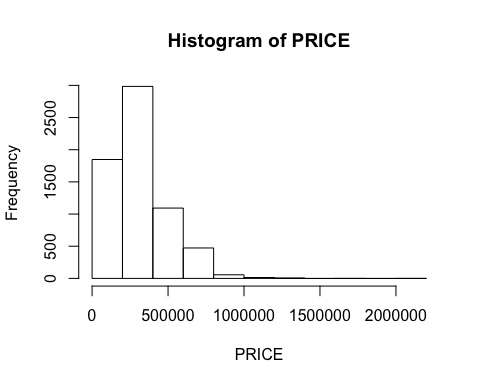
**new\_data** has been instantiated with the same covariates as before but it contains no values with "." associated. **new\_data** appears to be "cleaner" than data and thus we can proceed with describing the data more fully.

## ZIP: Look at boxplot and compare statistics  
boxplot(PRICE~ZIP)

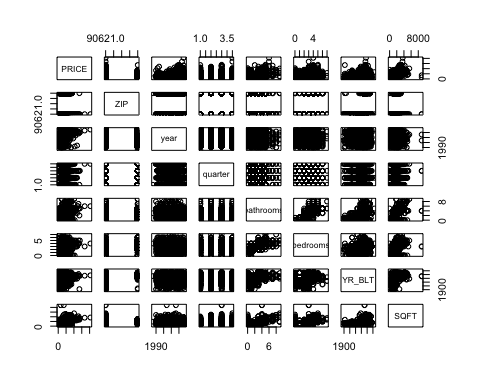


From the boxplot of the zip-code, we can see that for 90623 the average price of a house is higher than that that 90621 despite the latter having significantly more outliers.

hist(PRICE)



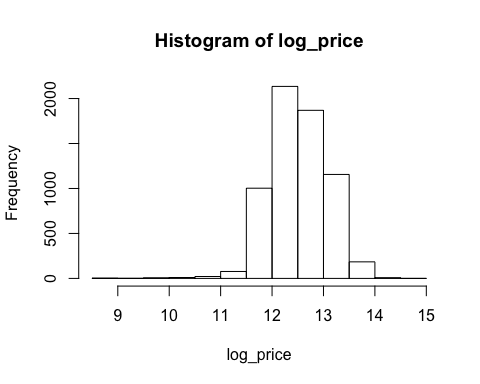
pairs(~PRICE + ZIP + year + quarter + bathrooms + bedrooms + YR\_BLT + SQFT)



From the PRICE histogram, we see the positive skew so we can transform the PRICE function with a log\_transformation. From the pairs function, we can also see that the other covariates are correlated quite strongly with price so when we create our model, there are strongly associated relationships between some of the covariates and the response variable. One thing to note here is that there might be some multicollinearity which may need correction via some variance inflation factor or a greedy algorithm which addresses it directly.

An alternative solution could be to use Principal Component Analysis (PCA) and find the corresponding eigenvalues that explain the most variance and then either create a regression using Partial Least Squares (PLS) or simply do a Principal Component Analysis. Despite the strength of these nonparametric models, if the multicollinearity isn't very high between the covariates, we can proceed with a multiple regression.

log\_price <- log(PRICE)  
hist(log\_price)



summary(log\_price)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 8.70 12.14 12.50 12.52 12.92 14.53

We can see here that the log transformation is appropriate for us to make better normality assumptions (despite the heavy tails on both ends). Given that year is not a continuous variable, it would make sense to make both year and YR\_BLT categorical variables. In addition, given that there are only two zip codes, we can change ZIP into a categorical variable as well. We can now proceed to fitting a model and model selection.

Given no clear insight as to which variables to choose from the list and given the relatively small number of covariates, we will create a fit with all of them and then perform backwards selection.

fit1<-lm(log\_price~ZIP+year+quarter+bathrooms+bedrooms+YR\_BLT+SQFT,data=new\_data)  
summary(fit1)

##   
## Call:  
## lm(formula = log\_price ~ ZIP + year + quarter + bathrooms + bedrooms +   
## YR\_BLT + SQFT, data = new\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.3654 -0.0722 0.0090 0.0955 1.3628   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.038e+01 3.516e-01 29.520 < 2e-16 \*\*\*  
## ZIP90623 1.682e-01 9.742e-03 17.268 < 2e-16 \*\*\*  
## year1990 1.287e+00 2.629e-01 4.896 1.00e-06 \*\*\*  
## year1991 1.217e+00 2.491e-01 4.886 1.05e-06 \*\*\*  
## year1992 1.183e+00 2.492e-01 4.746 2.12e-06 \*\*\*  
## year1993 1.089e+00 2.491e-01 4.373 1.25e-05 \*\*\*  
## year1994 1.075e+00 2.491e-01 4.314 1.63e-05 \*\*\*  
## year1995 1.004e+00 2.492e-01 4.028 5.70e-05 \*\*\*  
## year1996 9.836e-01 2.490e-01 3.950 7.90e-05 \*\*\*  
## year1997 1.026e+00 2.491e-01 4.120 3.83e-05 \*\*\*  
## year1998 1.126e+00 2.490e-01 4.523 6.19e-06 \*\*\*  
## year1999 1.213e+00 2.490e-01 4.871 1.14e-06 \*\*\*  
## year2000 1.301e+00 2.490e-01 5.225 1.80e-07 \*\*\*  
## year2001 1.398e+00 2.490e-01 5.615 2.05e-08 \*\*\*  
## year2002 1.557e+00 2.490e-01 6.253 4.30e-10 \*\*\*  
## year2003 1.773e+00 2.490e-01 7.123 1.17e-12 \*\*\*  
## year2004 2.031e+00 2.490e-01 8.157 4.10e-16 \*\*\*  
## year2005 2.186e+00 2.490e-01 8.780 < 2e-16 \*\*\*  
## year2006 2.276e+00 2.492e-01 9.134 < 2e-16 \*\*\*  
## year2007 2.176e+00 2.493e-01 8.728 < 2e-16 \*\*\*  
## year2008 1.936e+00 2.493e-01 7.768 9.24e-15 \*\*\*  
## year2009 1.825e+00 2.500e-01 7.300 3.22e-13 \*\*\*  
## quarter 1.415e-02 2.936e-03 4.821 1.46e-06 \*\*\*  
## bathrooms 3.700e-02 4.710e-03 7.855 4.68e-15 \*\*\*  
## bedrooms -2.535e-02 5.530e-03 -4.583 4.67e-06 \*\*\*  
## YR\_BLT1898 -3.107e-01 3.513e-01 -0.884 0.37656   
## YR\_BLT1902 -2.154e-01 2.883e-01 -0.747 0.45518   
## YR\_BLT1903 1.539e-01 2.873e-01 0.536 0.59217   
## YR\_BLT1904 -1.725e-01 2.656e-01 -0.650 0.51598   
## YR\_BLT1907 1.947e-01 3.514e-01 0.554 0.57943   
## YR\_BLT1910 -1.347e-01 2.637e-01 -0.511 0.60945   
## YR\_BLT1912 1.283e-01 2.781e-01 0.461 0.64457   
## YR\_BLT1913 -1.634e-01 3.044e-01 -0.537 0.59135   
## YR\_BLT1914 1.797e-01 3.514e-01 0.511 0.60917   
## YR\_BLT1916 -8.231e-01 2.779e-01 -2.962 0.00307 \*\*   
## YR\_BLT1917 8.487e-02 2.777e-01 0.306 0.75994   
## YR\_BLT1918 3.109e-01 3.044e-01 1.022 0.30706   
## YR\_BLT1919 -2.225e-02 2.778e-01 -0.080 0.93615   
## YR\_BLT1920 8.167e-02 2.585e-01 0.316 0.75205   
## YR\_BLT1921 -4.344e-02 2.579e-01 -0.168 0.86625   
## YR\_BLT1922 -9.677e-02 2.605e-01 -0.371 0.71030   
## YR\_BLT1923 -2.578e-04 2.547e-01 -0.001 0.99919   
## YR\_BLT1924 9.266e-03 2.562e-01 0.036 0.97115   
## YR\_BLT1925 7.414e-02 2.527e-01 0.293 0.76924   
## YR\_BLT1926 -1.406e-01 2.507e-01 -0.561 0.57492   
## YR\_BLT1927 -7.176e-02 2.513e-01 -0.286 0.77523   
## YR\_BLT1928 3.760e-02 2.498e-01 0.150 0.88038   
## YR\_BLT1929 2.640e-02 2.509e-01 0.105 0.91621   
## YR\_BLT1930 1.277e-02 2.552e-01 0.050 0.96009   
## YR\_BLT1933 2.107e-03 2.685e-01 0.008 0.99374   
## YR\_BLT1935 -5.748e-02 2.870e-01 -0.200 0.84127   
## YR\_BLT1937 7.368e-04 2.636e-01 0.003 0.99777   
## YR\_BLT1938 2.837e-02 2.567e-01 0.111 0.91200   
## YR\_BLT1939 -3.315e-01 2.658e-01 -1.247 0.21233   
## YR\_BLT1940 1.785e-02 2.533e-01 0.070 0.94380   
## YR\_BLT1941 -5.697e-02 2.580e-01 -0.221 0.82524   
## YR\_BLT1942 -8.830e-01 3.041e-01 -2.904 0.00370 \*\*   
## YR\_BLT1943 1.624e-01 3.513e-01 0.462 0.64398   
## YR\_BLT1944 -1.587e-01 2.722e-01 -0.583 0.55982   
## YR\_BLT1945 -4.345e-02 2.522e-01 -0.172 0.86322   
## YR\_BLT1946 -8.848e-04 2.494e-01 -0.004 0.99717   
## YR\_BLT1947 5.564e-02 2.491e-01 0.223 0.82326   
## YR\_BLT1948 -3.372e-02 2.496e-01 -0.135 0.89256   
## YR\_BLT1949 4.933e-03 2.512e-01 0.020 0.98433   
## YR\_BLT1950 1.796e-02 2.502e-01 0.072 0.94278   
## YR\_BLT1951 2.910e-02 2.534e-01 0.115 0.90857   
## YR\_BLT1952 9.937e-02 2.504e-01 0.397 0.69155   
## YR\_BLT1953 1.766e-03 2.527e-01 0.007 0.99442   
## YR\_BLT1954 -2.224e-02 2.499e-01 -0.089 0.92909   
## YR\_BLT1955 3.875e-03 2.510e-01 0.015 0.98768   
## YR\_BLT1956 2.257e-02 2.513e-01 0.090 0.92843   
## YR\_BLT1957 2.056e-01 2.492e-01 0.825 0.40924   
## YR\_BLT1958 2.126e-01 2.501e-01 0.850 0.39542   
## YR\_BLT1959 3.396e-02 2.496e-01 0.136 0.89178   
## YR\_BLT1960 1.081e-01 2.684e-01 0.403 0.68722   
## YR\_BLT1961 2.217e-02 2.525e-01 0.088 0.93004   
## YR\_BLT1962 -1.126e-04 2.609e-01 0.000 0.99966   
## YR\_BLT1963 7.003e-02 2.506e-01 0.279 0.77993   
## YR\_BLT1964 1.073e-01 2.513e-01 0.427 0.66940   
## YR\_BLT1965 6.497e-03 2.492e-01 0.026 0.97920   
## YR\_BLT1966 -4.473e-02 2.491e-01 -0.180 0.85750   
## YR\_BLT1967 1.529e-02 2.495e-01 0.061 0.95115   
## YR\_BLT1968 3.290e-02 2.495e-01 0.132 0.89509   
## YR\_BLT1969 3.011e-02 2.493e-01 0.121 0.90388   
## YR\_BLT1970 -1.222e-01 2.498e-01 -0.489 0.62477   
## YR\_BLT1971 -2.060e-01 2.490e-01 -0.827 0.40815   
## YR\_BLT1972 -4.599e-02 2.494e-01 -0.184 0.85369   
## YR\_BLT1973 -5.925e-02 2.490e-01 -0.238 0.81190   
## YR\_BLT1974 -1.771e-02 2.501e-01 -0.071 0.94356   
## YR\_BLT1975 2.075e-02 2.494e-01 0.083 0.93370   
## YR\_BLT1976 -1.114e-02 2.496e-01 -0.045 0.96440   
## YR\_BLT1977 6.292e-02 2.519e-01 0.250 0.80274   
## YR\_BLT1978 7.977e-02 2.505e-01 0.318 0.75017   
## YR\_BLT1979 -1.160e-02 2.623e-01 -0.044 0.96473   
## YR\_BLT1980 -1.329e-02 2.518e-01 -0.053 0.95791   
## YR\_BLT1981 -5.008e-02 2.590e-01 -0.193 0.84667   
## YR\_BLT1982 5.294e-02 2.689e-01 0.197 0.84395   
## YR\_BLT1983 2.470e-02 2.523e-01 0.098 0.92201   
## YR\_BLT1984 -1.530e-02 2.621e-01 -0.058 0.95345   
## YR\_BLT1985 2.974e-02 2.531e-01 0.117 0.90647   
## YR\_BLT1986 1.036e-01 2.776e-01 0.373 0.70907   
## YR\_BLT1987 -7.192e-03 2.507e-01 -0.029 0.97711   
## YR\_BLT1988 -1.405e-01 3.527e-01 -0.398 0.69039   
## YR\_BLT1989 1.970e-01 2.541e-01 0.775 0.43821   
## YR\_BLT1990 2.943e-02 2.500e-01 0.118 0.90630   
## YR\_BLT1993 -2.131e-02 2.783e-01 -0.077 0.93895   
## YR\_BLT1996 -5.812e-01 2.686e-01 -2.164 0.03052 \*   
## YR\_BLT2005 9.733e-01 3.526e-01 2.760 0.00580 \*\*   
## SQFT 3.248e-04 8.716e-06 37.265 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.248 on 6368 degrees of freedom  
## Multiple R-squared: 0.8031, Adjusted R-squared: 0.7998   
## F-statistic: 240.5 on 108 and 6368 DF, p-value: < 2.2e-16

fit2<-lm(log\_price~ZIP+year+quarter+bathrooms+bedrooms+SQFT,data=new\_data)  
summary(fit2)

##   
## Call:  
## lm(formula = log\_price ~ ZIP + year + quarter + bathrooms + bedrooms +   
## SQFT, data = new\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.4086 -0.0923 0.0139 0.1145 1.5131   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 10.2993573 0.2616722 39.360 < 2e-16 \*\*\*  
## ZIP90623 0.1138189 0.0071715 15.871 < 2e-16 \*\*\*  
## year1990 1.3069275 0.2755657 4.743 2.15e-06 \*\*\*  
## year1991 1.2331186 0.2617968 4.710 2.53e-06 \*\*\*  
## year1992 1.2093288 0.2618762 4.618 3.95e-06 \*\*\*  
## year1993 1.1084327 0.2617696 4.234 2.32e-05 \*\*\*  
## year1994 1.1022155 0.2617246 4.211 2.57e-05 \*\*\*  
## year1995 1.0328253 0.2617841 3.945 8.05e-05 \*\*\*  
## year1996 1.0073066 0.2616739 3.849 0.000120 \*\*\*  
## year1997 1.0439886 0.2616931 3.989 6.70e-05 \*\*\*  
## year1998 1.1382717 0.2616058 4.351 1.38e-05 \*\*\*  
## year1999 1.2318854 0.2616526 4.708 2.55e-06 \*\*\*  
## year2000 1.3214103 0.2616386 5.051 4.53e-07 \*\*\*  
## year2001 1.4141060 0.2616249 5.405 6.71e-08 \*\*\*  
## year2002 1.5810567 0.2616127 6.044 1.59e-09 \*\*\*  
## year2003 1.7964104 0.2615862 6.867 7.15e-12 \*\*\*  
## year2004 2.0569581 0.2616474 7.862 4.42e-15 \*\*\*  
## year2005 2.2009350 0.2616241 8.413 < 2e-16 \*\*\*  
## year2006 2.2978457 0.2618036 8.777 < 2e-16 \*\*\*  
## year2007 2.2102567 0.2619841 8.437 < 2e-16 \*\*\*  
## year2008 1.9656988 0.2619062 7.505 6.95e-14 \*\*\*  
## year2009 1.8522388 0.2626801 7.051 1.96e-12 \*\*\*  
## quarter 0.0158413 0.0030693 5.161 2.53e-07 \*\*\*  
## bathrooms 0.0340969 0.0040159 8.490 < 2e-16 \*\*\*  
## bedrooms -0.0185666 0.0053302 -3.483 0.000499 \*\*\*  
## SQFT 0.0003602 0.0000078 46.187 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2612 on 6451 degrees of freedom  
## Multiple R-squared: 0.7787, Adjusted R-squared: 0.7779   
## F-statistic: 908.1 on 25 and 6451 DF, p-value: < 2.2e-16

anova(fit1,fit2)

## Analysis of Variance Table  
##   
## Model 1: log\_price ~ ZIP + year + quarter + bathrooms + bedrooms + YR\_BLT +   
## SQFT  
## Model 2: log\_price ~ ZIP + year + quarter + bathrooms + bedrooms + SQFT  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 6368 391.76   
## 2 6451 440.23 -83 -48.47 9.4924 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

library(MASS)  
stepAIC(fit2)

## Start: AIC=-17362.83  
## log\_price ~ ZIP + year + quarter + bathrooms + bedrooms + SQFT  
##   
## Df Sum of Sq RSS AIC  
## <none> 440.23 -17362.8  
## - bedrooms 1 0.83 441.06 -17352.7  
## - quarter 1 1.82 442.05 -17338.1  
## - bathrooms 1 4.92 445.15 -17292.9  
## - ZIP 1 17.19 457.42 -17116.7  
## - SQFT 1 145.58 585.81 -15514.4  
## - year 20 1180.75 1620.98 -8960.1

##   
## Call:  
## lm(formula = log\_price ~ ZIP + year + quarter + bathrooms + bedrooms +   
## SQFT, data = new\_data)  
##   
## Coefficients:  
## (Intercept) ZIP90623 year1990 year1991 year1992   
## 10.2993573 0.1138189 1.3069275 1.2331186 1.2093288   
## year1993 year1994 year1995 year1996 year1997   
## 1.1084327 1.1022155 1.0328253 1.0073066 1.0439886   
## year1998 year1999 year2000 year2001 year2002   
## 1.1382717 1.2318854 1.3214103 1.4141060 1.5810567   
## year2003 year2004 year2005 year2006 year2007   
## 1.7964104 2.0569581 2.2009350 2.2978457 2.2102567   
## year2008 year2009 quarter bathrooms bedrooms   
## 1.9656988 1.8522388 0.0158413 0.0340969 -0.0185666   
## SQFT   
## 0.0003602

fit3<-lm(log\_price~ZIP+year+bathrooms+bedrooms+SQFT,data=new\_data)  
summary(fit3)

##   
## Call:  
## lm(formula = log\_price ~ ZIP + year + bathrooms + bedrooms +   
## SQFT, data = new\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.4040 -0.0933 0.0142 0.1152 1.5033   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.031e+01 2.622e-01 39.342 < 2e-16 \*\*\*  
## ZIP90623 1.145e-01 7.185e-03 15.930 < 2e-16 \*\*\*  
## year1990 1.351e+00 2.760e-01 4.896 1.00e-06 \*\*\*  
## year1991 1.258e+00 2.623e-01 4.795 1.66e-06 \*\*\*  
## year1992 1.235e+00 2.623e-01 4.706 2.58e-06 \*\*\*  
## year1993 1.135e+00 2.622e-01 4.327 1.53e-05 \*\*\*  
## year1994 1.128e+00 2.622e-01 4.301 1.73e-05 \*\*\*  
## year1995 1.059e+00 2.623e-01 4.039 5.42e-05 \*\*\*  
## year1996 1.031e+00 2.622e-01 3.934 8.44e-05 \*\*\*  
## year1997 1.070e+00 2.622e-01 4.082 4.51e-05 \*\*\*  
## year1998 1.163e+00 2.621e-01 4.436 9.30e-06 \*\*\*  
## year1999 1.257e+00 2.621e-01 4.796 1.65e-06 \*\*\*  
## year2000 1.346e+00 2.621e-01 5.137 2.87e-07 \*\*\*  
## year2001 1.439e+00 2.621e-01 5.488 4.21e-08 \*\*\*  
## year2002 1.604e+00 2.621e-01 6.122 9.80e-10 \*\*\*  
## year2003 1.820e+00 2.621e-01 6.946 4.14e-12 \*\*\*  
## year2004 2.081e+00 2.621e-01 7.940 2.37e-15 \*\*\*  
## year2005 2.224e+00 2.621e-01 8.485 < 2e-16 \*\*\*  
## year2006 2.322e+00 2.623e-01 8.852 < 2e-16 \*\*\*  
## year2007 2.230e+00 2.625e-01 8.497 < 2e-16 \*\*\*  
## year2008 1.993e+00 2.624e-01 7.594 3.53e-14 \*\*\*  
## year2009 1.862e+00 2.632e-01 7.076 1.64e-12 \*\*\*  
## bathrooms 3.375e-02 4.023e-03 8.390 < 2e-16 \*\*\*  
## bedrooms -1.884e-02 5.341e-03 -3.528 0.000422 \*\*\*  
## SQFT 3.613e-04 7.812e-06 46.252 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2618 on 6452 degrees of freedom  
## Multiple R-squared: 0.7778, Adjusted R-squared: 0.777   
## F-statistic: 941.1 on 24 and 6452 DF, p-value: < 2.2e-16

anova(fit2,fit3)

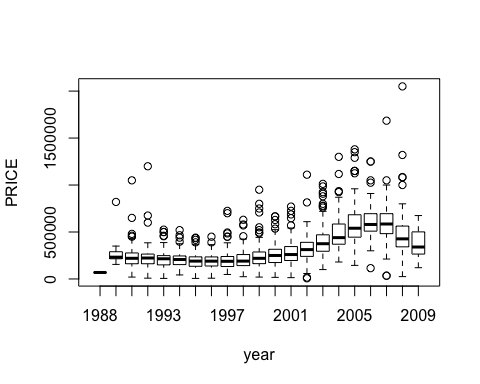
## Analysis of Variance Table  
##   
## Model 1: log\_price ~ ZIP + year + quarter + bathrooms + bedrooms + SQFT  
## Model 2: log\_price ~ ZIP + year + bathrooms + bedrooms + SQFT  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 6451 440.23   
## 2 6452 442.05 -1 -1.8178 26.638 2.527e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

After looking at fit1, it can be seen that YR\_BLT is not statistically significant and thus I removed it in favor for a tentatively better fit, fit2. The Multiple R-Squared value decreased by 3% but the residual standard error also decreased signficantly. After looking at the analysis of variance between the two, fit2 seems to be much better, having all statistically significant predictors and a relatively smaller error. After performing stepAIC, a greedy algorithm for backwards selection, one can notice that YRBLT doesn't contribute much to the fit and thus confirms the need for it to be removed.

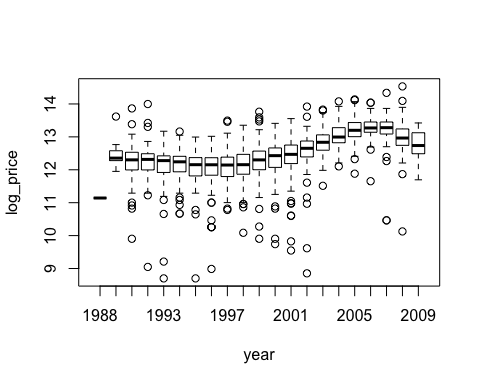
It would appear that the logprice seems to increase as the bathrooms and bedrooms increase upto a point, after which the price drops significantly, presumably because the lack of demand for such characteristics. The year built doesn't seem to have much effect on the logprice, nor does the quarter. The latter is slightly puzzling given that one hears so much about housing trends being cyclical during different seasons. It also appears that the zip code is a very statistically significant predictor of logprice. Given our assumptions about this data generating mechanism, it is statistically significant but predicting for other zip codes, it clearly doesn't have much predictive power.

One thing to note is that the data is not homoscedastic and due to the long tails on both sides of the log\_price distribution on the qq-normal plot, the model doesn't appear to be very normal. If we scale back after predicting for certain characteristics, we can assume normality.

plot(PRICE~year,data=new\_data)



plot(log\_price~year,data=new\_data)



It would appear that in both PRICE over time and logprice over time, house prices dipped around 1990 until 1998 (housing bubble) and then surged again until 2007 and then dropped once more (probably stock market crash and associated housing bubble).

It would appear that the statistically significant factors affecting house price are the ZIP code (to some degree) probably because of association of name, the year, the bathrooms and bedrooms (number), and the square feet.

The predicted function is fit3 (with the coefficients corresponding to them). After inputting for the regressors we can see that below:

fit3$coefficients

## (Intercept) ZIP90623 year1990 year1991 year1992   
## 10.3146134722 0.1144530020 1.3512770162 1.2576830796 1.2346921600   
## year1993 year1994 year1995 year1996 year1997   
## 1.1347301662 1.1277125564 1.0593156949 1.0313122178 1.0702437876   
## year1998 year1999 year2000 year2001 year2002   
## 1.1627255929 1.2572337681 1.3464567206 1.4385077046 1.6044743671   
## year2003 year2004 year2005 year2006 year2007   
## 1.8202000468 2.0812481093 2.2238755726 2.3216063780 2.2303460673   
## year2008 year2009 bathrooms bedrooms SQFT   
## 1.9925547152 1.8623720346 0.0337546656 -0.0188395708 0.0003613307

When you combine the regressor coefficients with fit3 and input certain parameters, that will output your log\_price after which you can simply exponentiate

The residual sum of squares is the error associated with the function and it's equivalent to **0.2618** which essentially is saying that given a certain prediction, the price is going to be log(0.2618) dollars off.

### FUTURE DIRECTIONS:

As mentioned earlier, a non-parametric approach using PCR (principal component regression) would be much more efficient and better fitting, considering that the covariance matrix for the covariates given is not so tedious to compute the second and third eigenvalues. It'll also do a better job of handling multicollinearity and also at explaining variance. This would be a better approach for the future (as well as using regsubsets from the leaps package for backwards and exhaustive model selection).