Week06: Housing Data Analysis

Kurt Stoneburner

7/7/2020

a. Explain why you chose to remove data points from your 'clean' dataset. As a currently active house hunter the primary factors that affect price are:

1. Size
2. Location
3. Bedrooms
4. Bathrooms
5. House Quality

Other factors that may have a lesser effect on price

- Age of the home - There may be a premium based on the age of the home.
- Sale Warning - May indicate other issues with the property or repairs, that may deflate the selling price.
- Zoning variation - Is the location zoned differently than use? This implies a redevlopment scenario, where an older home on a larger lot is replaced with a higher density structure(s). Although it is a possibility I looked at the relationship between zoning and present use. A unique list of zones was generated. The current_zoning column was converted to a numeric vector based on the index value in the unique zone list. The numeric zones were correlated with the present_use column.

The resulting value:

```
## [1] 0.02921141
```

portrays no significant link between these variables.

All other categories unrelated categories were removed.

One other note on the data set. The three bathroom variables have been summed into a new variable bath_total. Using the following weighted values: bath_full_count + (bath_half_count *.5) + (bath_3qtr_count *.75)

The bath_total is more in line with current listing standards and makes the relationship of bathroom count to the overall price easier to interpret.

b. Create two variables; 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. Explain the basis for your additional predictor selections.

1. The an increase in total living space should increase the sale price since structure costs increase with size.

2. Building Grade denotes the overall quality of the structure which should directly affect price
3. Bedrooms and Bathrooms directly reflect the needs of the American buyer which assumes individual space demands a higher premium than the overall shared space of the dwelling.
4. The age of the house should affect price. It is expected that an increasing house age will have negative effect of Sales Price.

Correlation confirms these assumptions.

```
cor(housing_df)[1,]
```

```
##               Sale.Price                     zip5           building_grade
##               1.00000000               0.06014866               0.39122909
## square_feet_total_living                 bedrooms               year_built
##               0.45458758               0.22546748               0.24267127
##                sq_ft_lot               bath_total                house_age
##               0.11981223               0.35449263              -0.22681160
```

Based on the correlation, variables with a correlation of greater than .20 are good candidates to work with. I was genuinely surprised at the correlation value of bathrooms to sale price. A linear model can built with these variables.

```
salePrice_base_lm <-  lm(Sale.Price ~ sq_ft_lot, data=housing_df)
salePrice_house_age_lm <-    lm(Sale.Price ~ square_feet_total_living + building_grade + bedrooms + bath
```

Taking the age of the house into account feels like an important element of the model. I'm struggling with which is a better metric house_age, or year built. Year built has a slightly better Adjusted R squared and F score, which indicates it would be a slightly better metric. However, house age appears to affect the model as expected, the older the house the more negative impact on price. Since the parameter seems to better follow expectations (and the impact is small) i'm sticking with house age as the appropriate metric.

c. Execute a summary() function on two variables defined in the previous step to compare the model results. What are the R2 and Adjusted R2 statistics? Explain what these results tell you about the overall model. Did the inclusion of the additional predictors help explain any large variations found in Sale Price?

```
##
## Call:
## lm(formula = Sale.Price ~ sq_ft_lot, data = housing_df)
##
## 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 ~ square_feet_total_living + building_grade +
##     bedrooms + bath_total + house_age, data = housing_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1711510  -119020   -44977    41651  3939584
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                70526.468  33287.633   2.119   0.0341 *
## square_feet_total_living     152.772      6.233  24.508  < 2e-16 ***
## building_grade             32315.730   4461.197   7.244 4.61e-13 ***
## bedrooms                  -10303.193   4663.195  -2.209   0.0272 *
## bath_total                  2880.287   7207.962   0.400   0.6895
## house_age                  -1950.880    204.420  -9.543  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 357200 on 12859 degrees of freedom
## Multiple R-squared:   0.22,  Adjusted R-squared:  0.2197
## F-statistic: 725.5 on 5 and 12859 DF,  p-value: < 2.2e-16
```

Price by Lot size Model - Multiple R-squared: 0.01435, Adjusted R-squared: 0.01428 Price by Living Space ~ House Age - Multiple R-squared: 0.22, Adjusted R-squared: 0.2197

The Adjusted R-squared values significantly improved using additional predictors. The adjusted R-squared value for using Lot size as a predictor had barely any significance on predicted price. As indicated by the correlation calculation the Building Grade accounted for the largest variability in predicting price. I am a little surprised that bedrooms and bath_total have a negative coefficient using the year_built mode. The House age model reflects negative coefficents for house age and bedrooms. A negative relationship with house_age makes sense, and older house would be expected to have some negative pricing affect.

d. Considering the parameters of the multiple regression model you have created. What are the standardized betas for each parameter and what do the values indicate?

	square_feet_total_living	building_grade	bedrooms	bath_total	house_age	year_built
house_age	0.373945132	0.087315978	-0.022322776	0.004951635	-0.084478319	

For every increase in standard deviation of a given variable, the sale price will change by a stand deviation of sale price * Beta. Assuming the other variables remain constant.

For Example: Looking at the standard deviations for the variables:

```
##         variable      std_dev
## 1     Sale Price 4.043811e+05
## 2    sqft Living 9.898176e+02
## 3 Building Grade 1.092624e+00
## 4       bedrooms 8.761273e-01
## 5      bathrooms 6.951902e-01
## 6      house age 1.751079e+01
```

```
Predicted Sale Price increases **$151,216** per **989.817** increase in square_feet_total_living
  - 989.817 std deviation of square_feet_total_living.
  - 151,216 = 404,381.10 (sale price standard deviation) \* .373945132 (square foot living Beta)

Predicted Sale Price increases **$35308.93** per **1.09** increase in building_grade
  - 1.09 std deviation of building_grade
  - 35308.93 = 404,381.10 (sale price standard deviation) * 0.087315978 (building_grade Beta)

Predicted Sale Price decreases **$9026.90** per **.88** increase in bedrooms
  - .88 std deviation of building_grade
  - $9026.90 = 404,381.10 (sale price standard deviation) * -0.022322776 (bedrooms Beta)

Predicted Sale Price increases **$2002.35** per **.7** increase in bathrooms
  - .7 std deviation of building_grade
  - $2002.35 = 404,381.10 (sale price standard deviation) * 0.004951635 (bathrooms Beta)

Predicted Sale Price decreases **$-34,212.12** per **17.5** increase in house_age
  - 17.5 std deviation of building_grade
  - $-34,212.12 = 404,381.10 (sale price standard deviation) * -0.084478319 (house_age Beta)
```

All of these predicted values assume the other variables remain constant. This also indicates the variable bath_total has $< .001$ influence on the predicted sale price. Making it a good candidate for elimination.

e. Calculate the confidence intervals for the parameters in your model and explain what the results indicate. The 95% confidence intervals for Sales Price looks like this.

```
##   lowerBoundary     mean upperBoundary lowerValue upperValue
## 1      653749.9 660737.7      667725.6   6987.825    1328463
```

The lower and upper bounds indicate a range from the mean where 95% of the Sales Price values lie. 95% of the Sales Price data lies between $6,987.83 and $1,328,463. (If this data set was from the Bay Area, $1.3m would not be an outlier).

Looking at the 95% confidence interval for Square Feet Living Space

```
##   lowerBoundary     mean upperBoundary lowerValue upperValue
## 1      2522.402 2539.506      2556.611   17.10434   5096.117
```

The range of values for Square feet living is notable that the lower bounds is 17sqft which is an impractical size. There may be data entry errors or properties that are lot only (0 Square feet living). These are outliers that should be taken into consideration. The upper value of 5096 sqft indicates there very well may be some very large properties for sale. Considering wealth and income distribution issues in America, these may be a reflection of wealth distribution, therefore i'm not sure if should be considered outliers for the purposes of the model.

Looking at the 95% confidence interval for Building Grade

```
##   lowerBoundary    mean upperBoundary lowerValue upperValue
## 1      8.221539 8.24042      8.259301 0.01888086   16.49972
```

A lower value of near zero likely indicates there are properties sold without buildings. I suspect this would align with properties with 0sq Ft. An upper bound of 16, indicates there are definitely some fancy homes in the top 5% of building grades.

Looking at the 95% confidence interval for bedrooms

```
##   lowerBoundary     mean upperBoundary lowerValue upperValue
## 1      3.463523 3.478663      3.493803 0.01513974   6.972466
```

Looking at the 95% confidence interval for bathrooms

```
##   lowerBoundary     mean upperBoundary lowerValue upperValue
## 1      2.463599 2.475612      2.487625 0.01201309   4.963237
```

Looking at the 95% confidence interval for House Age

```
##   lowerBoundary     mean upperBoundary lowerValue upperValue
## 1      17.81144 18.11403      18.41662  0.3025916   36.53065
```

95% of the homes are 0 - 36 years old.

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.
Looking at the analysis of variance

```
## Analysis of Variance Table
##
## Model 1: Sale.Price ~ sq_ft_lot
## Model 2: Sale.Price ~ square_feet_total_living + building_grade + bedrooms +
##     bath_total + house_age
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1  12863 2.0734e+15
## 2  12859 1.6407e+15  4 4.3266e+14 847.74 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The model shows an improved F score over the original model

```
casewise_df <- housing_df
casewise_df$cooks.distance <- cooks.distance(salePrice_house_age_lm)
casewise_df$leverage <- hatvalues(salePrice_house_age_lm)
casewise_df$covariance.ratios <- covratio(salePrice_house_age_lm)
```

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.

```
large_residuals <- rstandard(salePrice_house_age_lm) > 2 | rstandard(salePrice_house_age_lm) < -2
```

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.

i. Use the appropriate function to show the sum of large residuals. There are 325 large residuals, which represents 0.02526234 of the data. In a normal distribution, having more than 5% of the data attributed to large residuals is an indicator of undue influence on the model.

```
sum(large_residuals)
```

```
## [1] 325
```

```
sum(large_residuals) / nrow(housing_df)
```

```
## [1] 0.02526234
```

j. Which specific variables have large residuals (only cases that evaluate as TRUE)? Large_residuals is a logistical vector which can be used to select data that evaluates to TRUE. This method is used to create a data frame containing only the large outliers.

```
large_residuals_df <- housing_df[large_residuals,c(
  "Sale.Price",
  "building_grade",
  "square_feet_total_living",
  "bedrooms",
  "bath_total",
  "house_age")]
```

Each parameter can be compared to the upper and lower bounds of the confidence interval to identify the specific outlier values.

```
## [1] "Sales.Price - 245 Total outliers"
```

```
## [1] "              7 contain values less than 6988"
```

```
## [1] "              238 contain values more than 1328463"
```

```
## [1] "Building Grade - 0 Total outliers"
```

```
## [1] "Square Foot Living - 54 Total outliers"
```

```
## [1] "              54 contain values more than 5096"
```

```
## [1] "Bedrooms - 9 Total outliers"
```

```
## [1] "              5 contain values less than 0.02"
```

```
## [1] "              4 contain values more than 7"
```

```
## [1] "Bathrooms - 26 Total outliers"
```

```
## [1] "              3 contain values less than 0.01"
```

```
## [1] "              23 contain values more than 5"
```

```
## [1] "House Age - 259 Total outliers"
```

```
## [1] "              225 contain values less than 0.3"
```

```
## [1] "              34 contain values more than 37"
```

The predominant outliers appear to be large expensive homes.

k. Investigate further by calculating the leverage, cooks distance, and covariance rations. Comment on all cases that are problematics. There are no data with a Cooks Distance greater than one, therefore none of the data is having an outsized effect on the model. `{ r cooks_analysis, echo=TRUE}` `sum(casewise_df$cooks.distance > 1)`

Checking leverage indicates a number of outliers. Nearly 5% of the data is attributed to having double the average leverage. Along with 2.5% attributed to triple the average leverage. Taking into account no data has a Cooks Distance greater than one, and the abnormal leverage is limited to 5% the model is still valid.

```
## [1] "Number of cases with more than double the leverage:"
```

```
## [1] 726
```

```
## [1] "Representative percentage of the data"
```

```
## [1] 0.05643218
```

```
## [1] "Number of cases with more than triple the leverage:"
```

```
## [1] 329
```

```
## [1] "Representative percentage of the data"
```

```
## [1] 0.02557326
```

There are 472 cases that deviate significantly above the the covariance ratios, and 251 that deviate below. Again considering Cooks distance and relatively small sample size this should not have an outsized effect on the model.

```
#### Checking if the covariance ratios deviate 3 times from the average leverage
upperCov <- 1 + 3*(k + 1)/n
lowerCov <- 1 - 3*(k + 1)/n

sum(casewise_df$covariance.ratios > upperCov)
```

```
## [1] 472
```

```
sum(casewise_df$covariance.ratios < lowerCov)
```

```
## [1] 251
```

```
dwt(salePrice_house_age_lm)
```

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       0.7333545     0.5332807       0
##  Alternative hypothesis: rho != 0
```

The Durbin Watson statistic is less than 1 with a p of zero indicating the independence of errors condition is not met.

m. Perform the necessary calculations to assess the assumption of no multicollinearity and state if the condition is met or not. vif() is used to test for no multicollinarity. The vif value and tolerance generate no cause for concern. The mean indicates there may be some bias in the data. The textbook states if the value substantially deviates from 1 there may be bias. There is no reference or scale for substantial, so i'm not sure how to intrepret a value of 2.

```
### VIF > 10 is cause for Concern
vif(salePrice_house_age_lm)
```

```
## square_feet_total_living          building_grade                bedrooms
##                 3.838144                2.395488                1.682874
##               bath_total              house_age
##                 2.531526                1.291840
```

```
### VIF Tolerance < 0.2 cause for concern
1/vif(salePrice_house_age_lm)
```

```
## square_feet_total_living          building_grade                bedrooms
##                0.2605426               0.4174514               0.5942215
##               bath_total              house_age
##                0.3950187               0.7740899
```

```
### VIF mean substantially greater than 1 may be biased
mean(vif(salePrice_house_age_lm))
```

```
## [1] 2.347974
```

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. Plotting the linear model generates four plots.

Plot#1 - Residuals vs fitted. The expected outcome is a random distribution of datapoints. The plotted residuals tend to cluster near the fitted values indicating there is an issue linearity and randomness in the data.

Plot#2 - Q-Q plot that show standardized residuals in deviations from normal. There is a significant portion of the model that deviates from the norm.

Plot#3 Square root of the standardized residuals vs fitted values - Values closer to the line indicate a normal distribution. As with other plots there are significant indications that housing in not normally distributed.

Plot#4 Standardized residuals vs leverage. The values indicate no there are no cases that exert undue influence over the model.

a Histogram of the standardized residuals indicates a leptokurtic distribution.

`{ r plot_hist, echo=TRUE } plot(salePrice_house_age_lm) hist(salePrice_house_age_lm)`
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?

This regression model appears to be relatively unbiased. Cooke's distance indicates there are no data points exerting undue influence although there are abnormalities for the leverage and covariance ratios. The data appears to indicate that housing prices are not normally distributed.