Real Estate Sales - Linear vs Tree Regression

David Oniani

December 16, 2019

Introduction

This data set is from Appendix C, Table 7, of the book, Applied Linear Statistical Models, Fifth Edition, by Kutner, Nachtsheim, Neter and Li. It has data on 522 residential real estate transactions in a midwestern city in 2002. The variables are:

- Sales price in dollars
- Finished area in square feet
- Number of bedrooms
- Number of bathrooms
- Air conditioning (1=Yes, 0=No)
- Garage size (# of cars)
- Pool (1=Yes, 0=No)
- Year built - values range from 1885 to 1998
- Quality (1=High, 2=Medium, 3=Low)
- Style (Qualitative indicator of architectural style - Definition of specific levels not available)
- Lot size in square feet
- Highway (1=Adjacent to highway, 0=Not adjacent)

The response variable is sales price. We are interested in determining which variables can best predict sales price.

```
# Read in the data

realest = read.table(file="./APPENC07.txt",
                     header=F,
                     col.names=c("ID", "sales.price", "sqfeet", "bedrms",
                                 "bathrms", "aircond", "garsize", "pool",
                                 "yrbuilt", "quality", "style", "lotsize",
                                 "highway"))
```

Final Linear Model

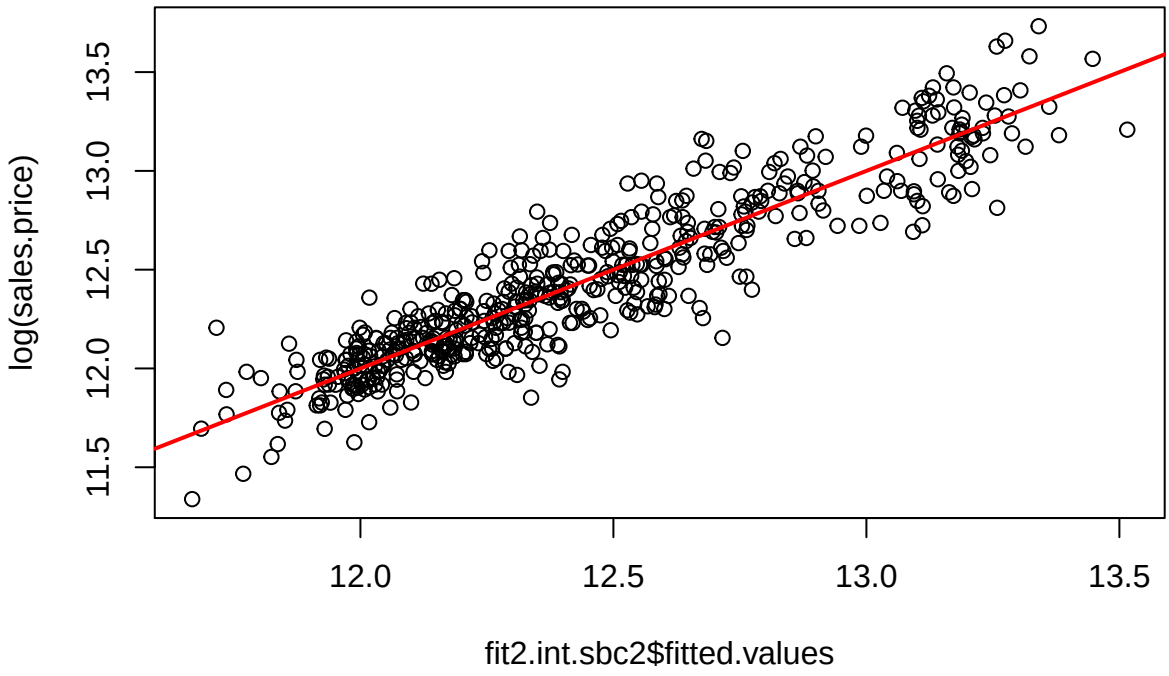
```
realest$log.sqft = log (realest$sqfeet)
realest$log.sqft.c = realest$log.sqft - mean (realest$log.sqft)

fit2.int.sbc2 = lm (log(sales.price) ~ log.sqft.c + bathrms + yrbuilt +
    as.factor(quality) + lotsize + log.sqft.c:as.factor(quality),
    data = realest)
summary (fit2.int.sbc2)
```

```
##
## Call:
## lm(formula = log(sales.price) ~ log.sqft.c + bathrms + yrbuilt +
##     as.factor(quality) + lotsize + log.sqft.c:as.factor(quality),
##     data = realest)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.56085 -0.10131  0.00043  0.10365  0.49055
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     4.507e+00  1.108e+00   4.067 5.52e-05 ***
## log.sqft.c                      2.532e-01  1.067e-01   2.372  0.01805 *
## bathrms                         3.849e-02  1.180e-02   3.262  0.00118 **
## yrbuilt                         4.173e-03  5.584e-04   7.474 3.39e-13 ***
## as.factor(quality)2            -5.259e-01  4.936e-02 -10.653  < 2e-16 ***
## as.factor(quality)3            -6.253e-01  5.668e-02 -11.032  < 2e-16 ***
## lotsize                         4.883e-06  6.716e-07   7.271 1.34e-12 ***
## log.sqft.c:as.factor(quality)2  5.962e-01  1.121e-01   5.320 1.55e-07 ***
## log.sqft.c:as.factor(quality)3  3.552e-01  1.308e-01   2.715  0.00684 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1689 on 513 degrees of freedom
## Multiple R-squared:  0.8492, Adjusted R-squared:  0.8468
## F-statistic:   361 on 8 and 513 DF,  p-value: < 2.2e-16
```

```r
plot (log(sales.price) ~ fit2.int.sbc2$fitted.values, data=realest)
abline (0, 1, col='red', lwd=2)
```



Interaction Plot

Interaction effect between square feet and quality:

```r
par (mfrow=c(1,1))
# Function to categorize a continuous variable into its quartiles
```

```r
categorize = function (x) {
  quartiles = summary (x) [c(2, 3, 5)]
  result = rep ("Q1", length (x))
  result [(quartiles[1] < x) & (x <= quartiles [2])] = "Q2"
  result [(quartiles[2] < x) & (x <= quartiles [3])] = "Q3"
  result [quartiles[3] < x] = "Q4"
  return (result)
}

# Interaction plots using the ggplot and dplyr packages

library (ggplot2)
library (dplyr)
```

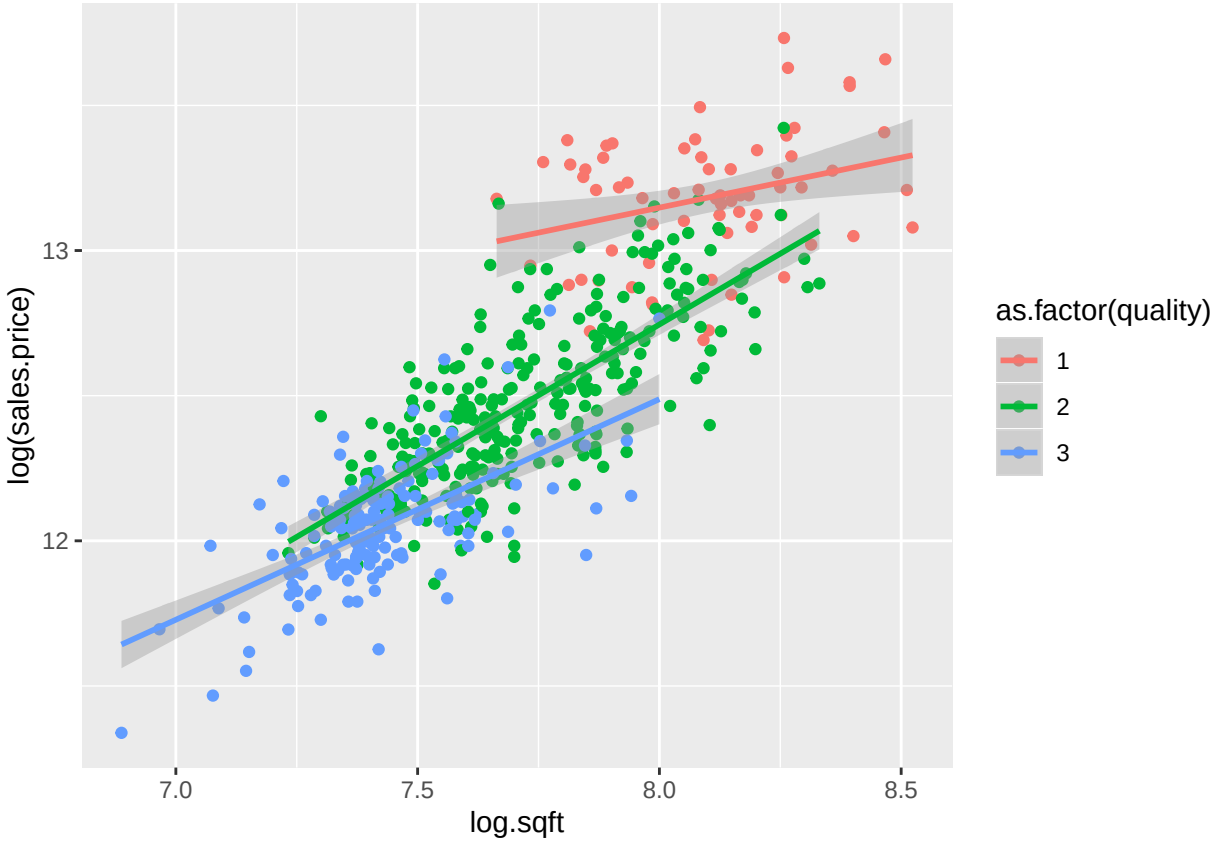
```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
# Plot Log sales price vs Log sqfeet by quality level
# Note the use of the "with" function to avoid having to specify
# each column name with "realest$" in front of it

with (realest,
      qplot (x=log.sqft, y=log(sales.price), color=as.factor (quality)) +
        geom_smooth (method="lm"))
```



This interaction plot shows that the relationship between log(sales price) and log(sqfeet) is stronger (steeper) for medium and low quality homes than it is for high quality homes (quality=1).

The final model has adjusted $R^2 = 0.85$, which means that 85% of the variation on log sales price is explained by the model. The residual standard error is 0.167 log dollars.

Tree Regression

For comparison to the final linear model, we fit a tree regression using log (sales.price). Use all of the predictor variables in the data set, not just the ones selected in the final linear regression model.

```
par(mfrow=c(1, 1))

library (rpart)
realest_tree = rpart(log(sales.price) ~ sqfeet + bedrms + bathrms + aircond + garsize +
                                        pool + yrbuilt + quality + style + lotsize + highway,
                                        realest)
```

Find the tree with the smallest xerror:

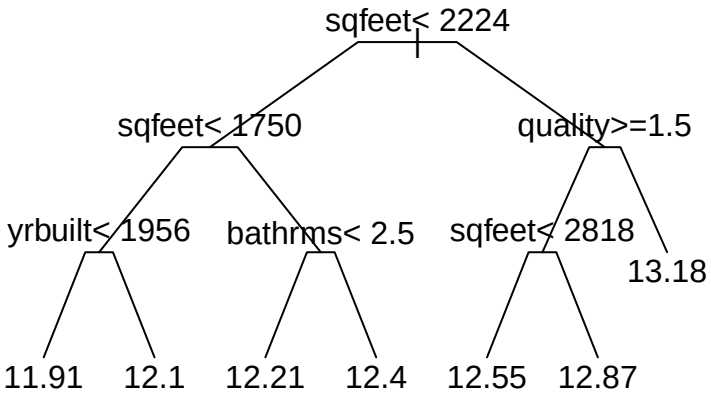
```
xerror_min = which.min(realest_tree$cptable[, "xerror"])
xerror_min
```

```
## 7
## 7
```

Plot and interpret the tree:

```
par (mfrow=c(1, 1))

# Plot the tree (click "Zoom")
plot(realest_tree, uniform=TRUE, margin=0.25, branch=0.25, compress=TRUE)
text(realest_tree)
```



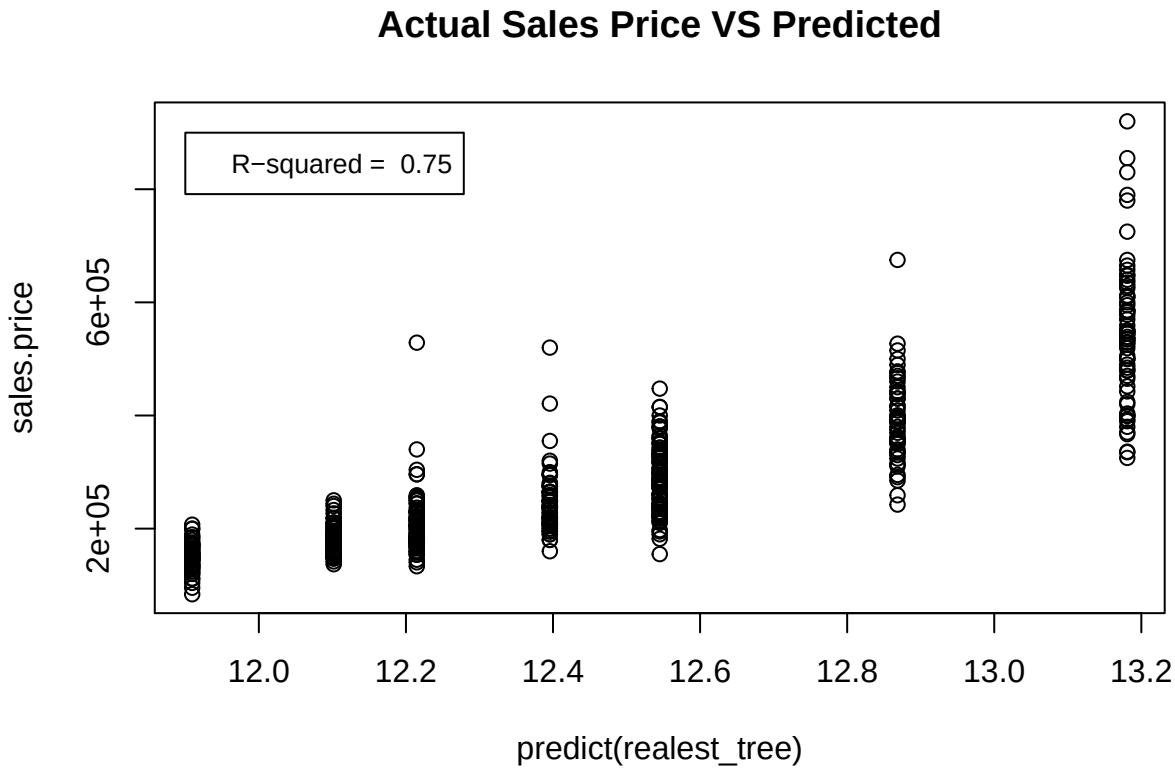
This tree says that sqfeet is the single most important predictor. Houses have the highest sales price when sqfeet $>= 2224$ and quality < 1.5. That group of houses has an average sales.price of 13.18 log(dollars). In this case, all other predictors except for sqfeet and quality do not matter.

The houses with the lowest sales.price are ones with sqfeet < 2224, sqfeet < 1750, and yrbuilt < 1956. Those houses have an average sales price of 11.91 log(dollars). In this case, all other predictors except for sqfeet and yearbuilt do not matter. As a side note, sqfeet < 2224 and sqfeet < 1750 can be written as sqfeet < 1750.

bathrms matters when sqfeet < 2224 and sqfeet $>= 1750$. Houses with more bathrooms have a higher price, on average.

Plot Log (sales price) vs. predicted values and interpret the plot. Report R^2 for the tree model.

```
with(realest, plot(predict(realest_tree), sales.price, main="Actual Sales Price VS Predicted"))
abline(0, 1, col="red")
rsq = cor(predict(realest_tree), realest$sales.price)^2
legend(11.9, 9e5, c(paste("R-squared = ", round (rsq, 2))), cex=0.8)
```



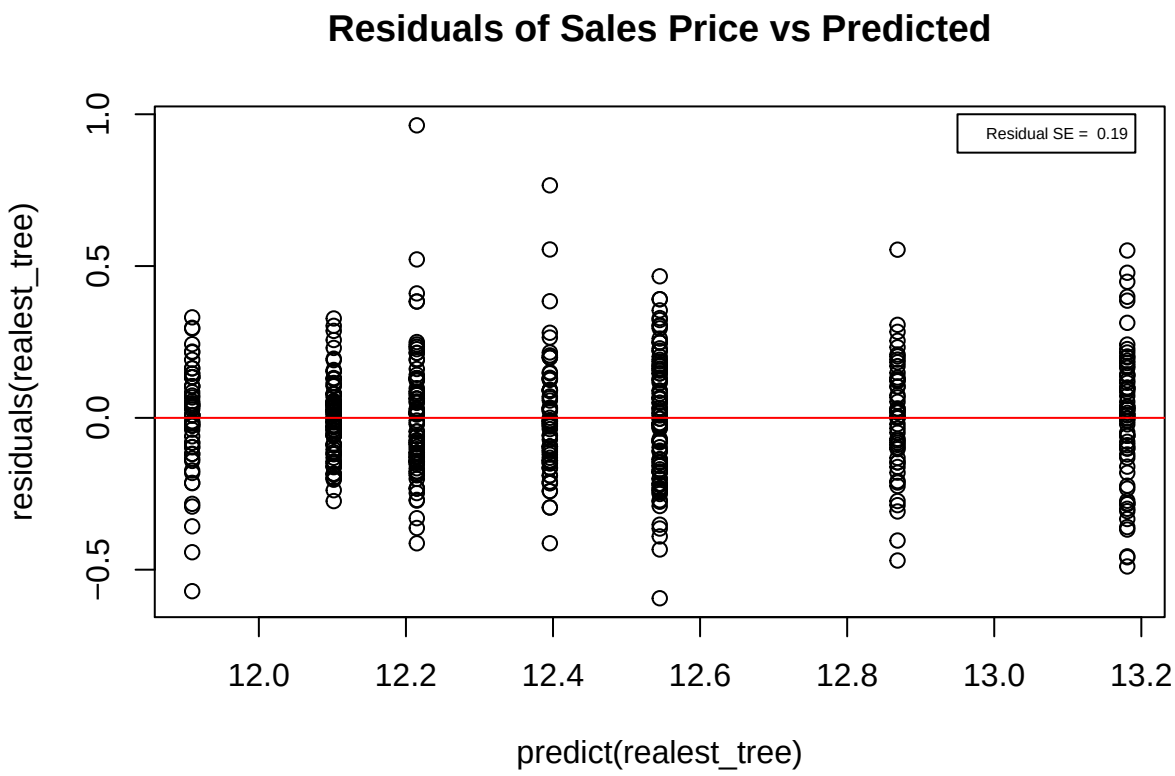

```
# R-squared
rsq
```

```
## [1] 0.7544671
```

The R-squared value for the the tree model is 0.75. Therefore, the relative error for the training data is 0.25. Also, 75% of the variation in log(sales.price) is explained by this tree regression model. The plot does seem to follow the linear trend, which should be the case by definition. A few outliers might be present. That said, the number of these points can be considered negligible as compared to the total number of points.

Plot residuals vs predicted values and interpret the plot. Report the residual standard error for the tree model.

```
plot(predict(realest_tree), residuals(realest_tree), main="Residuals of Sales Price vs Predicted")
abline(0, 0, col="red")
resid_se = sd(residuals(realest_tree))
legend(12.95, 1, c(paste("Residual SE = ", round(resid_se, 2))), cex=0.5)
```




```
# Residual standard error
resid_se
```

```
## [1] 0.1890254
```

The residual standard error is approximately 0.19 log(dollars) which is approximately 1.2 dollars (when exponentiated). This number can be considered almost negligible as compared to the real estate prices. The horizontal line splits the plot in approximately two halves which tells us that the errors are spread equally.

Compare the linear regression and tree regression results. In particular,

- Which predictors were most significant in each model?

The linear model suggested that quality was the most significant regressor variable. On the other hand, the tree regression suggested that sqfeet was the most significant regressor variable.

- Which predictors were selected for both models, vs. for one model or the other?

bathrms, quality, and yrbuilt were selected for both models. **If we also count log-transformed variables**, then variable sqfeet was selected for both models as well.

Variable lotsize was selected for the linear model, but not for the tree regression model. **If we do not count log-transformed variables**, then variable log.sqft.c was present in the linear model, but not in the tree regression model and variable sqfeet was present in the tree regression model, but not in the linear model.

- Does the interaction effect that was found in the linear regression show up in the tree regression model? Explain how that interaction effect is similar and/or different in the tree model vs. the linear regression?

Yes, it does (**given that we count the log-transformed variable**).

In the linear model, the interaction plot showed that the relationship between log(sales price) and log(sqfeet) is stronger (steeper) for medium and low quality homes than it is for high quality homes (quality=1).

Similar to the linear model, in the tree regression model, the relationship between sales price and sqfeet is stronger for medium and low quality homes (quality >= 1.5).

- For each predictor variable, describe whether the trend (direction of effect) is similar or different between the tree model and the linear regression model?

Tree regression:

- sales price and quality have a negative relationship
- sales price and sqfeet have a positive relationship
- sales price and bathrms have a positive relationship
- sales price and yrbuilt have a positive relationship

Linear regression:

- sales price and quality have a negative relationship
- sales price and sqfeet have a positive relationship
- sales price and bathrms have a positive relationship
- sales price and yrbuilt have a positive relationship
- sales price and lotsize have a positive relationship