**Stat 587: lab 11 – self assessment**

We will evaluate the relationships between home sales price and assessed values for 84 homes sold in one neighborhood of Tampa, Florida between 2008 and 2009. The data are in tamsales1.csv on the class web site. Sales price and assessed values are in units of 1000$, so $200,000 is 200 in the data set.

Each row of data is for the sale of one house. The three variables are the sales price, the assessed value of the land, and the assessed value of the improvements (buildings, e.g. the house, garage, or garden shed, or an in-ground pool). The goal is to develop a model to predict sales price for a home about to be put on the market.

1. Graphically evaluate whether a straight line is appropriate to model the relationship between sales price and the assessed value of the land.
2. Fit a model predicting the sales price from the assessed value of the land and the assessed value of the improvements. What is the prediction equation?
3. How precise are predictions of sales prices?
4. Comment on the appropriateness of the regression assumptions.

The next two questions are based on the model that uses LAND, IMPROVE and their product LAND\*IMPROVE to predict PRICE. If using JMP, please turn off Center Polynomials or create the product yourself.

1. Is including the cross product: LAND\*IMPROVE useful? Answer this three ways:
	1. Is the slope estimate for LAND\*IMPROVE significantly different from zero?

Use a t test and report the T statistic and p-value

* 1. Is the slope estimate for LAND\*IMPROVE significantly different from zero?

Use a model comparison test and report the F statistic and p-value

* 1. How much does adding LAND\*IMPROVE improve the prediction of sales price?

**Answers:**

1. Yes there it is, but there is moderate amount of scatter. See SalePrice vs Land plot in the scatterplot matrix.



1. Price = -103 + 2.01 Land + 1.78 Improve
2. rMSE = estimated sd of observations around regression line = 146 (note: I rounded a bit)
3. The residual vs predicted value plot doesn’t look great.



I see unequal variance (larger prediction sd for more expensive houses). Although a bit hard to see because of the number of overlapping dots, I see lack of fit for many houses with small predicted values (all those residuals are > 0, not spread above and below the 0 line).

Note: If nothing else, the unequal variance issue suggests a problem with the ‘how accurate are the predictions’ numbers. Ask if you don’t see why.

1. a) There is very strong evidence that the coefficient for Land\*Improve is not 0:

T = 4.39, p < 0.0001

 b) There is very strong evidence that the coefficient for Land\*Improve is not 0

 F = 19.24, p < 0.0001 (Note, as expected F = t2)

c) Adding LAND\*IMPROVE reduces the rMSE to 132 (rounded a bit).

Note: That is a 10% increase in precision. That could be a useful addition, but the error is still large.