**ISOM2600 InClass Case Exercise (Class 6)**

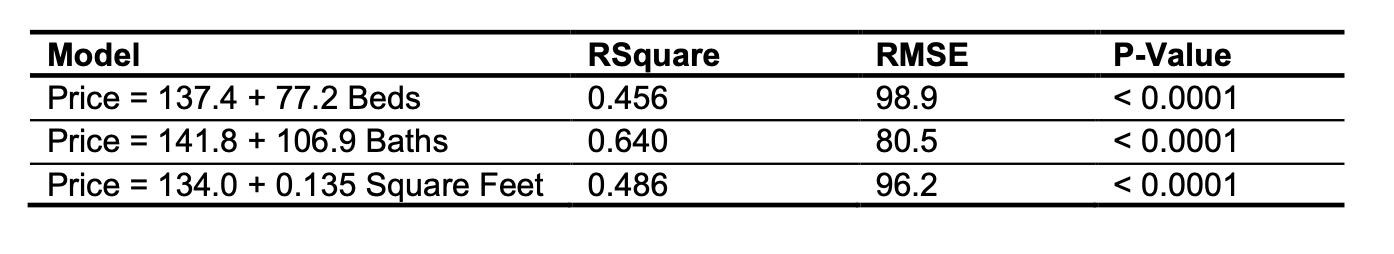
A real estate company that manages properties around a ski resort in the United States wishes to improve its methods for pricing homes. Data is readily available on a number of measures, including size of the home and property, location, age of the house, and a strength-of-market indicator. After determining which factors relate to the selling prices of homes located in and around the ski resort, develop a model to predict housing prices.

The data set contains information about 45 residential properties sold during a recent 12-month period. The variables in the data set are:

* Price: Selling price of the property ($1,000)
* Beds: Number of bedrooms in the house
* Baths: Number of bathrooms in the house
* Square Feet: Size of the house in square feet
* Miles to Resort: Miles from the property to the downtown resort area
* Miles to Base: Miles from the property to the base of the ski resort’s mountain
* Acres: Lot size in number of acres
* Cars: Number of cars that will fit into the garage
* Years Old: Age of the house, in years, at the time it was listed
* DoM: Number of days the house was on the market before it sold

This problem involves one response variable, the selling price of the home, and various potential predictors of selling price. Three of the predictor variables measure, directly or indirectly, the size of the house: square feet, number of bedrooms and number of bathrooms. Simple regression models were developed for each of these predictors.

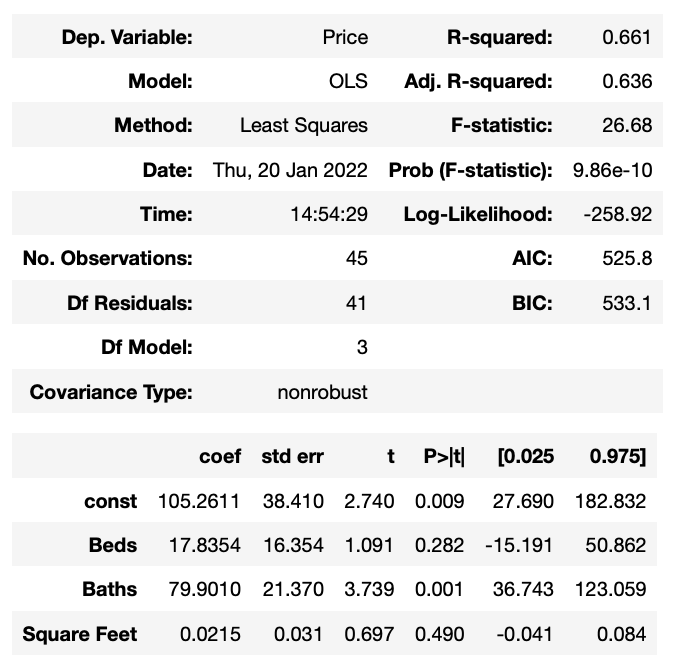
Below are summaries for each of the models:



The models indicate that, based on the single-predictor models, the house price increases by:

* $77,200 for each bedroom.
* $106,900 for each bathroom.
* $135 for each square foot.

Multiple regression allows us to understand the simultaneous influence of the three predictor variables on the selling price.:

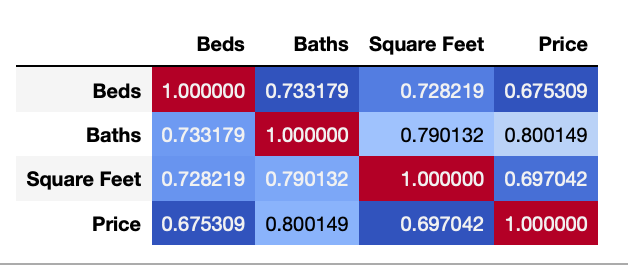


Comparing the multiple regression model to the three simple regression models reveals that the coefficients have changed.

**Question-1: Which of the following is false about the changes of the coefficient from simple linear to multiple regression?**

1. **$17,835 per bedroom in the multiple regression model, down from $77,200.**
2. **$79,900 per bathroom (formerly $106,900).**
3. **$21/square foot in multiple regression, but $135/square foot in simple regression.**
4. **The estimated intercept in multiple linear regression is larger than those in simple linear regression**

In addition, the significance of each of the predictors has changed. Two of the predictors that are statistically significant by themselves in simple regression models (Beds and Square Feet) are no longer significant when used in conjunction with other predictors. The possible reason for this is the *multicollinearity*. The correlations matrix of them is given below:



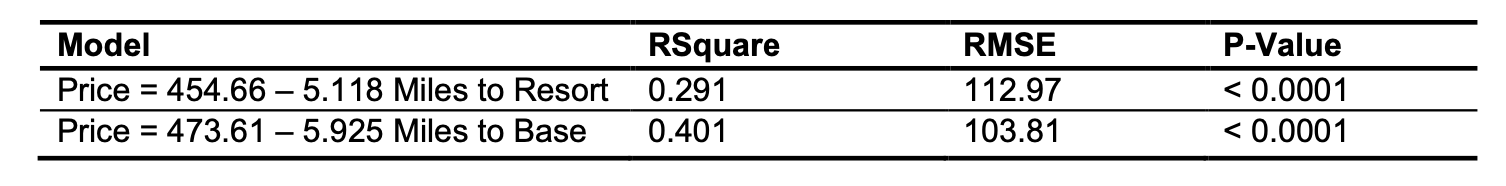
**Question-2: Which of the following is not the cause of multicollinearity?**

1. **The number of bathrooms is highly correlated with the price.**
2. **Adding a bedroom to a house tends to adding its square feet.**
3. **Adding a bathroom to a house tends to adding its square feet.**
4. **A larger house tends to have more bedrooms/bathrooms.**

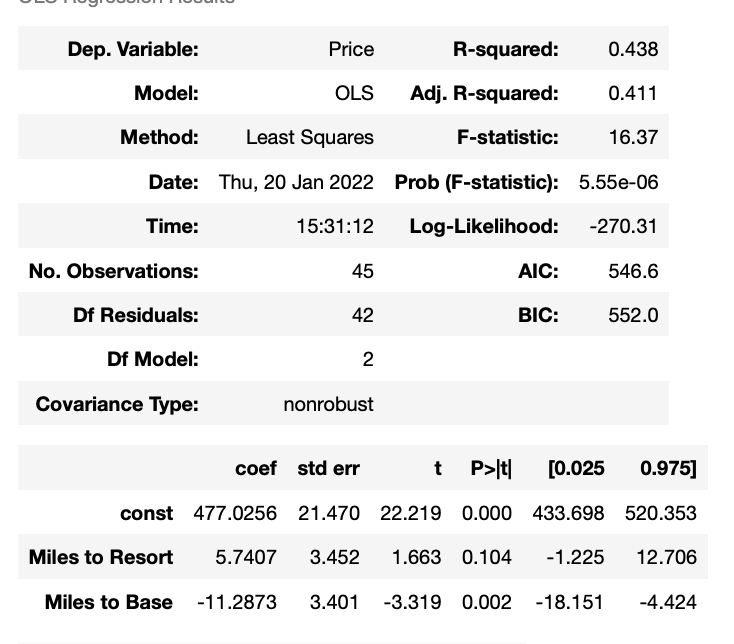
Multicollinearity might also be anticipated between the two location measures: miles to the mountain base and miles to the downtown resort area. Their correlations suggest that there is indeed reason for concern, since there is high correlation between the two predictors (0.948).



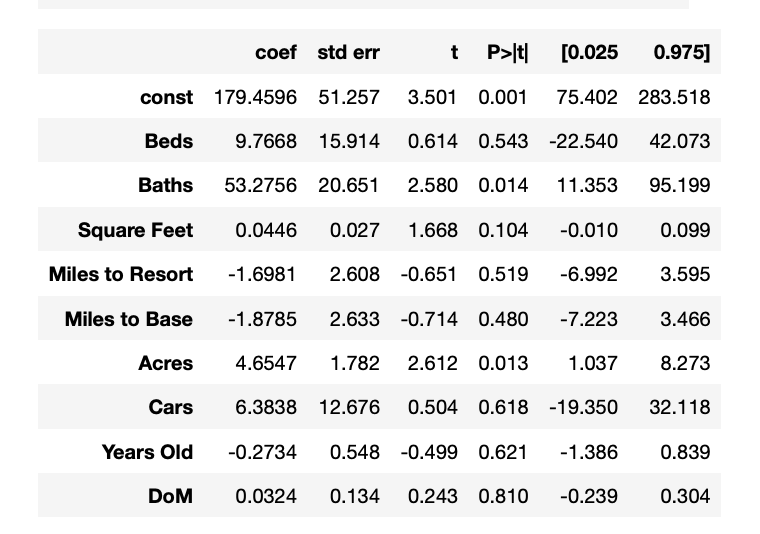
In the simple regression models for the location variables, both predictors are highly significant and the coefficients are negative.



The overall regression model involving the two location variables is highly significant, with a p- value of < 0.0001 for the F Ratio. Yet only Miles to Base is a significant predictor of Price (at the 0.05 level). In addition, the coefficients have changed dramatically. The coefficient for Miles to Resort is now positive!



We’ve found indications of multicollinearity with subsets of predictors, but let’s take a look at the multiple regression model using all of the predictors.



**Question-3: Which variables are significant at 5% in the full regression model?**

1. **Beds and Baths**
2. **Baths and Acres**
3. **Baths and Square Feet**
4. **Beds and Acres**

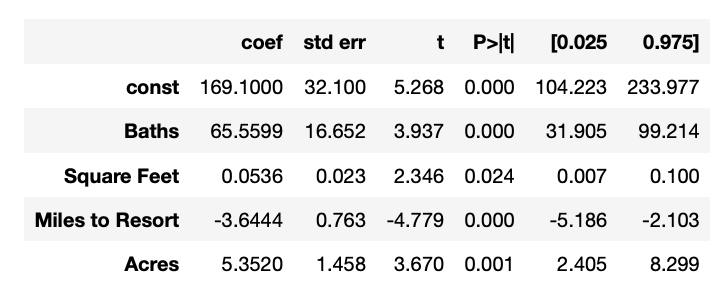
Are the other predictor variables not statistically significant due to multicollinearity? A measure of the severity of the multicollinearity is the variance inflation factor, or VIF.



**Question-4: Which variables result in multicollinearity according to VIF? And what treatment can we do?**

1. **Miles to Base and Miles to Base; We may delete either one since these two variables have similar meaning/measure.**
2. **Miles to Base and Miles to Base; We may create a ratio variable between them as it utilize both variables into one.**
3. **Beds and Square feet; We may delete either one since these two variables have similar meaning/measure.**
4. **Beds and Square feet; We may create a ratio variable between them as it utilize both variables into one.**

After handling the multicollinearity problem and further remove non-significant variables one by one, the final model is reached below:



The final model estimates selling prices at:

* $53.64 per square foot.
* $65,560 per bathroom.
* $5,352 per acre.
* $3,644 less for each mile away from the resort.

**Question-5: Suppose a house with the following features, what is the predicted price from the final model?**

* **2 baths**
* **1600 sqft**
* **15 miles apart from resort**
* **0.5 acre**

1. **$334,000**
2. **$165,000**
3. **$443,000**
4. **$230,000**

The model tells us not to worry too much about garage capacity, age of the home and days on the market when it comes to estimating a house’s selling price in this particular market. Being closer to the downtown resort area and mountain raises the selling price, and larger lots demand a higher price.