Lab 7

STA 216 F19

In this lab, we examine a random sample of 200 homes taken from public records in Saratoga County, New York. (This is a larger version of the dataset we first looked at on Handout 10.) The data is given on Blackboard in the file “lab 7 – real estate.xlsx”. The variables of interest in this lab include:

|  |  |
| --- | --- |
| SAS variable name | Description |
| Price | Price of the home in dollars |
| Living\_area | Square feet of “living space” in the house |
| Fireplaces | Number of fireplaces in the house |

Suppose you are a contractor who specializes in building fireplaces and chimneys, and you want to know how much value a fireplace adds to a house in this region.

1. After reading in the data, use PROC FREQ to summarize how many houses have the different numbers of fireplaces. Copy and paste the summary table.
2. The vast majority of houses have either zero or one fireplace, so we will make a dichotomized version of this variable.
   1. Copy and paste the code that adds a dummy variable named Dfp to the dataset that takes the value 1 if a house has at least one fireplace and the value 0 if a house doesn’t have any fireplaces.
   2. Add labels and formats to Dfp so the SAS output will be interpretable for people unfamiliar with your definition of Dfp. Copy and paste results from PROC FREQ for this variable that verifies that Dfp is defined correctly and has a proper label and formats.
3. Summarize the relationship between price and whether or not the house has a fireplace.
   1. Copy and paste a table giving the mean and standard deviation of price for the houses in both groups.
   2. Copy and paste side-by-side boxplots.
   3. Report the difference in the average prices of houses (to the nearest dollar) depending on whether or not they have a fireplace.
4. Let’s consider the square feet too and investigate whether it tends to differ based on whether the house has a fireplace.
   1. Copy and paste a table giving the mean and standard deviation of square feet for the houses in both groups.
   2. Copy and paste side-by-side boxplots.
5. Copy and paste a scatterplot that examines how square feet affects the price, with different types of points for houses that have/don’t have a fireplace. Also, use the Snipping Tool to draw two additional points on this plot, showing the location of the (square feet, price) means for houses with and without fireplaces.
6. Does the number in 3c represent the value of adding a fireplace to a house? It should seem awfully large. Keeping in mind what you found in #4 and 5, explain why the 3c number is probably larger than a fireplace’s value.
7. Let’s use a regression model to estimate the value of fireplace, while taking the square feet into account. Use PROC REG to model price in terms of square feet and Dfp.
   1. Write the equation of the least squares regression line, using the variable names and (hat) where appropriate.
   2. Copy and paste the portion of the SAS output that you used to get this line.
   3. Plug in values for Dfp to find the equations of the regression lines   
      [Price(hat) = \_\_\_\_\_ + \_\_\_\_\*sqft] for
      1. Houses without fireplaces
      2. Houses with fireplaces
   4. Show a scatterplot with these two regression lines using the LINEPARM statement (twice).
   5. Interpret the slope of living\_area in the context of the problem.
   6. Interpret the slope of Dfp in the context of the problem.
   7. Which of these estimates the value of adding a fireplace?
   8. Explain why this is a better estimate than the number in 3c.
   9. Is there statistically significant evidence that at least one of living\_area or Dfp is useful in predicting price? Report the test statistic and p-value associated with this conclusion.
   10. Is there statistically significant evidence that the price depends on Dfp when square feet is in the model? Report the test statistic and p-value associated with this conclusion.
   11. Report a 95% confidence interval for the slope of square feet.
8. Copy and paste a scatterplot of price vs. square feet with separate regression lines (with different slopes) for houses with and without a fireplace. Use PROC SGPLOT with the REG statement.
9. Use WHERE statements to obtain the equations of these regression lines   
   [Price(hat) = \_\_\_\_\_ + \_\_\_\_\*sqft]
   1. For houses without a fireplace
   2. For houses with a fireplace
10. Use a multiple regression model that contains both of these lines.
    1. Copy and paste the code that creates the necessary interaction term.
    2. Write the equation of the least squares regression line with three explanatory variables, using the variable names and (hat) where appropriate.
    3. Plug in values for Dfp to find the equations of the regression lines   
       [Price(hat) = \_\_\_\_\_ + \_\_\_\_\*sqft] for
       1. Houses without fireplaces
       2. Houses with fireplaces

(Note: these should be the same as #11)

* 1. Interpret the slope of Dfp\*sqft.

1. Use the root MSE and adjusted R-squared values to determine which of the following three models best predicts price:  
   - the simple linear regression model with square feet  
   - the model that has separate parallel lines by Dp  
   - the model that has lines with different slopes by Dp
   1. Report all of the root MSE and adj R-sq values.
   2. Which model is the best? Explain why.