FIN 510 Big Data Analytics in Finance

Lab 7: Linear Regression

Due on 09/25/2021

Predicting Home Sale Prices

The file ameshousing.csv contains sale prices and home characteristics of 300 homes located in Ames, Iowa from 2006 to 2010.

In lecture 7, we ran a simple linear regression of sale prices (SalePrice) on the above ground living area (Gr_Liv_Area), where Gr_Liv_Area is a continuous numeric variable. In this assignment, we will transform Gr_Liv_Area to a categorical variable and then fit a simple linear regression of SalePrice on bins of Gr_Liv_Area.

0) Load the packages and suppress scientific notation

Use library() to load ggplot2. Use options(scipen=999) to suppress scientific notation.

1) Create a data frame

Load the data and save the result in a data frame named housing.df. Return the first six rows and column names using head() and names(), respectively.

2) Identify thresholds of four bins that have an equal number of observations

To create four bins that have an equal number of observations in each bin, we need to use function quantile() to identify five thresholds: the minimum value (0 quartile), the first quartile, the second quartile (median), and the third quartile, and the maximum value (fourth quartile).

The first parameter in function quantile() is the variable's name, housing.df$Gr_Liv_Area. The second parameter is a vector of probabilities. For example, use 0.5 to represent the second quartile, because 50% of data lies below this value.

Hint: use probs = c(0,0.25,0.5,0.75,1) to return the minimum value (0 quartile), the first quartile, the second quartile (median), and the third quartile, and the maximum value (fourth quartile).

Save the result as bins.

3) Bin Gr_Liv_Area

To transform a continuous variable into a categorical variable, use function .bincode() to bin Gr_Liv_Area according to the thresholds found in question 2.

The first parameter in function .bincode() is the variable’s name, housing.df$Gr_Liv_Area. The second parameter specifies the cut points, which should be bins without names. Use include.lowest = TRUE as the third parameter to include the lowest value, 334, in the first bin.

Hint: unname(bins) removes the names of a named vector. For example, bins <- unname(bins) returns the values only.

Save the binned above ground living area as Gr_Liv_Area_bin in housing.df. It takes a value of 1 if Gr_Liv_Area is in [334, 952], a value of 2 if Gr_Liv_Area is in (952, 1135], a value of 3 if Gr_Liv_Area is in (1135, 1337.25], and a value of 4 if Gr_Liv_Area is in (1337.25, 1500].

Use head() and str() to return the first six values and the data type of Gr_Liv_Area_bin.

4) Fit a linear regression model with an ordinal categorial predictor

Use lm() to fit a linear regression model with an ordinal categorical predictor, Gr_Liv_Area_bin.

Gr_Liv_Area_bin should be treated as an ordinal categorical variable, where the order matters but not the difference between values. For example, the difference between Gr_Liv_Area categories [334, 952] and (952, 1135] does not have the same meaning as the difference between Gr_Liv_Area categories (952, 1135] and (1135, 1337.25].

Hint: as.factor() converts Gr_Liv_Area_bin into a categorical variable in the regression. Thus, use formular SalePrice ~ as.factor(Gr_Liv_Area_bin) in lm().

Save the regression model as lm and return summary(lm).

5) Calculate the mean of SalePrice for each value in Gr_Liv_Area_bin

Use function tapply() to calculate SalePrice for each value in Gr_Liv_Area_bin. The first parameter in the function tapply() is the housing.df$SalePrice and the second parameter represents the group variable, housing.df$Gr_Liv_Area_bin. Use mean as the third parameter to specify the function.

What is the mean of SalePrice for homes where Gr_Liv_Area_bin is equal to 2 (category 2)? What is the mean of SalePrice for homes where Gr_Liv_Area_bin is equal to 1 (category 1)? What is the group mean difference between category 2 and category 1?

What is the mean of SalePrice for homes where Gr_Liv_Area_bin is equal to 3 (category 3)? What is the group mean difference between category 3 and category 1?

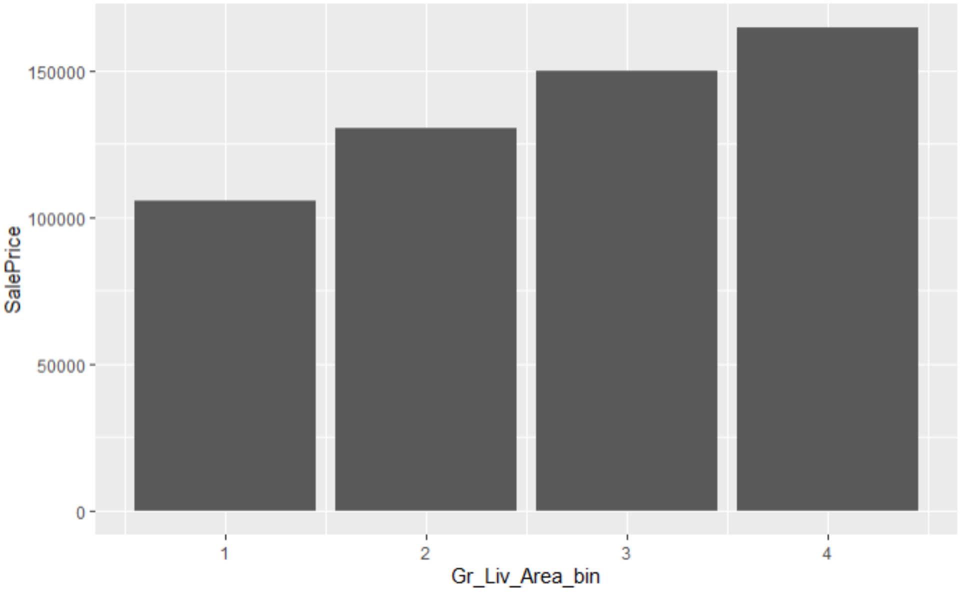
What is the mean of SalePrice for homes where Gr_Liv_Area_bin is equal to 4 (category 4)? What is the group mean difference between category 4 and category 1?

Notice that the regression coefficients are group mean differences relative to the omitted category (category 1).

6) Plot the mean of SalePrice for each value in Gr_Liv_Area_bin

Use ggplot() to plot the mean of SalePrice for each value in Gr_Liv_Area_bin.

The plot looks like the following graph.


SalePrice
Gr_Liv_Area_bin