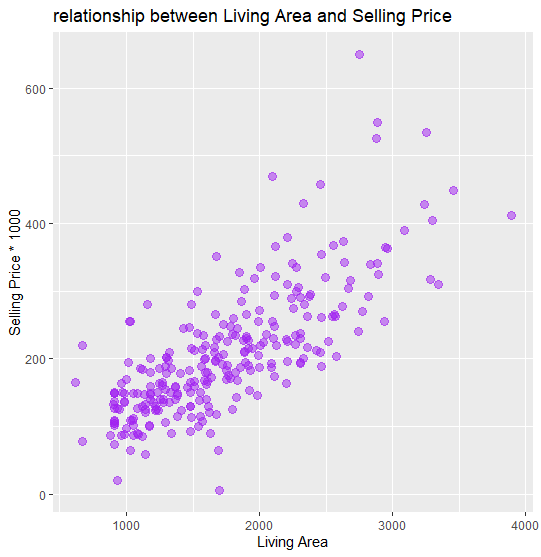
Project Report:

1. Construct a scatterplot of the selling price (Y) against the living area (X). Comment on the association between the two variables.

A: the scatterplot of the selling price (Y) against the living area (X) has been show below:



In general, the value of Y (Selling Price) is on its increase when the value of X (Living Area) is on its increase.

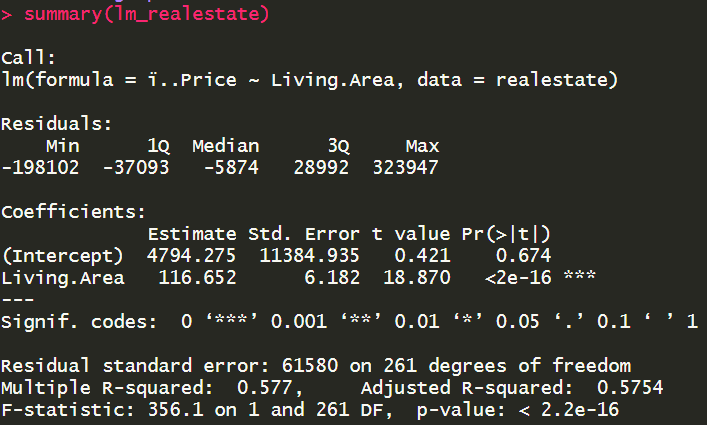
1. Explain why the selling price is selected to be the response variable.

A: According to the background material, we are discussing about other factors which could determine the price at which a house should be listed; in other words, we want to find other possible exploratory or predictor variables on which our selling price can depend. Therefore, the variation of selling price is followed by the variation of the other factors. Therefore, we set “selling price” as our response variable.

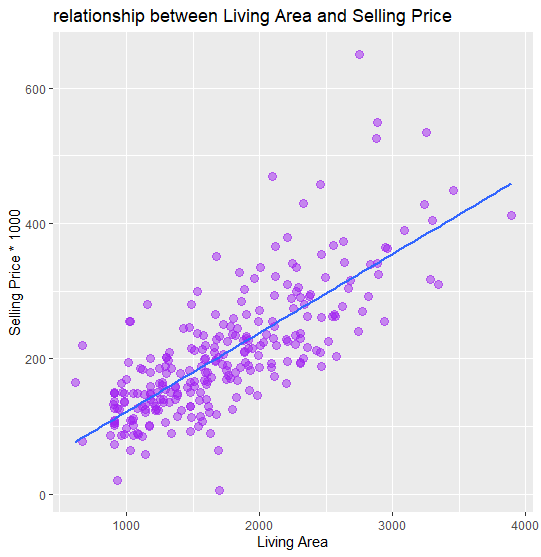
1. Regress the selling price on the living area. State the estimated regression function; and superimpose it on the scatterplot in part (1)

A: the regression function is stated below:

Y = 116.652X + 4794.275

(X: Living Area, Y: Selling Price)

The regression line on the scatterplot is showed below:

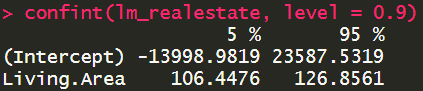


1. Report the values of MSE and R2 in the context of the problem.

A: The value of MSE is : 3791575129

 The value of R2 is : 0.5770478

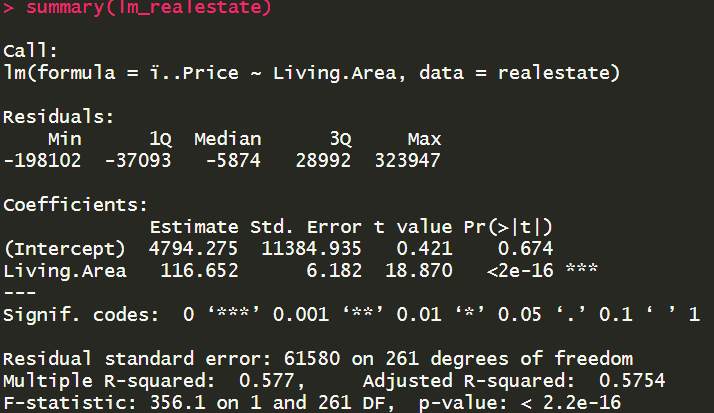
1. Obtain a 90% confidence interval for β1, and interpret it in the context of the problem.

A: (106.4476, 126.8561)

1. Test whether β1 is significantly different from zero at the 0.10 level of significance. State the alternatives, the value of t-test statistic, p-value, and your conclusion.

A: Ha: β1 is not equal to zero at the 0.10 level of significance.

t-test statistic value: 18.870

 p-value: less than 2e-16 (3-asterisk significance)

Conclusion: we should reject H0 and accept Ha, that is, β1 is not equal to zero at the 0.10 level of significance.

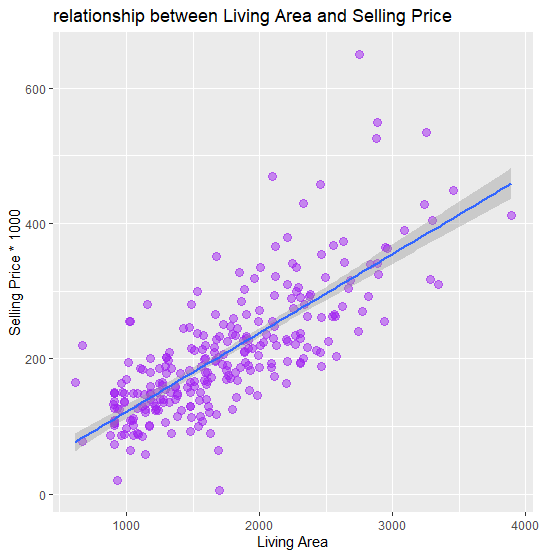
1. Predict the selling price for a home with 1850 square feet. Obtain a corresponding 90% prediction interval.

A: The predicted price of a home with 1850 squared feet is 220600.2

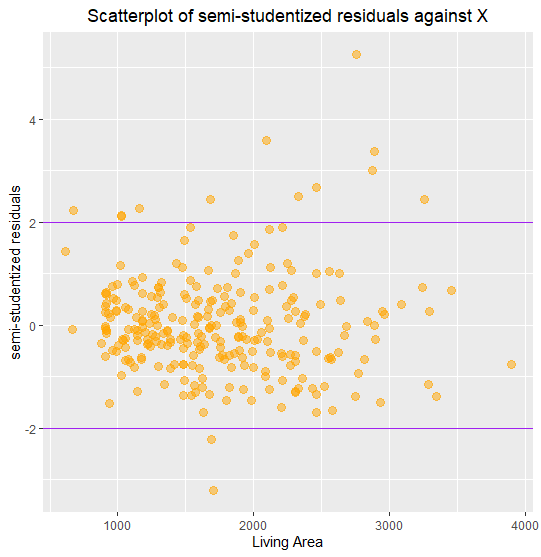


1. Superimpose the (pointwise) 90% prediction band on the scatterplot in part (1)

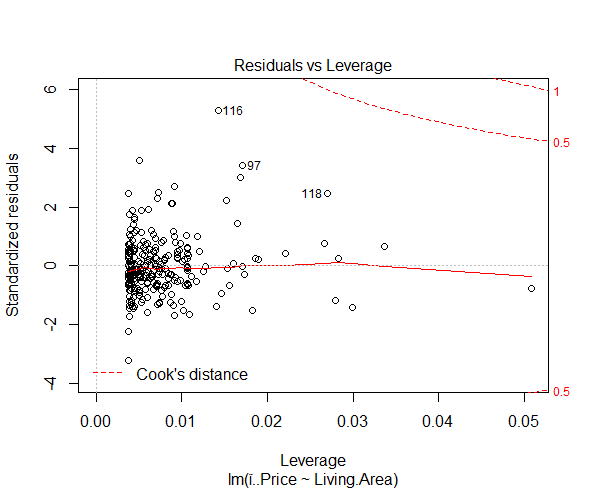
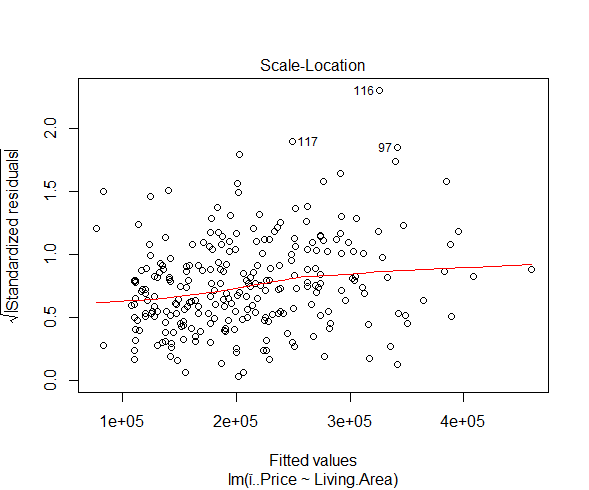
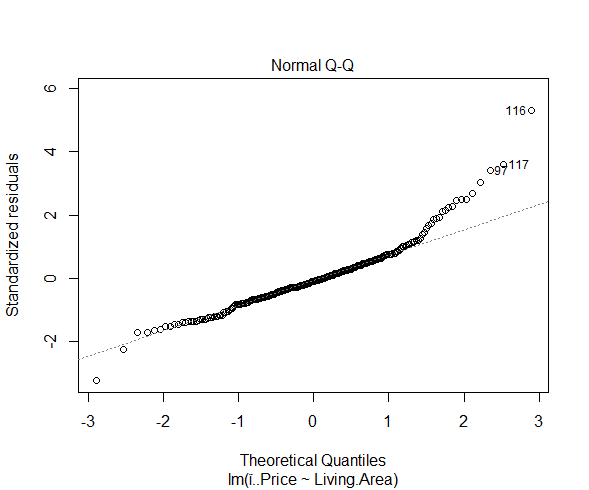
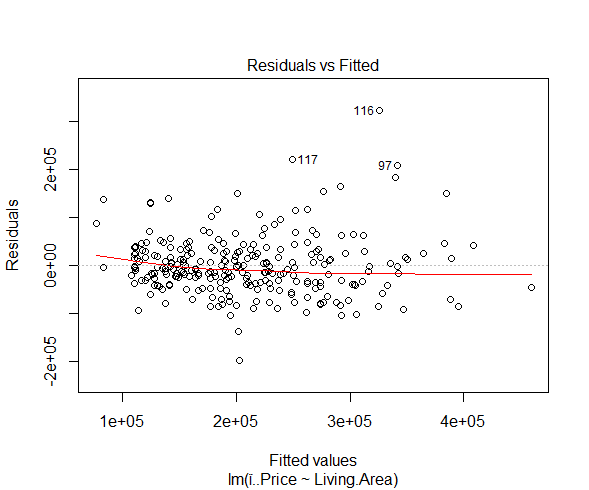
A: the plot has been showed below with 90% prediction band.



1. Construct a scatterplot of semi-studentized residuals against X and a normal probability plot of semi-studentized residuals. Use the two plots to perform model diagnostics.

A: (1) the scatterplot of semi-studentized residuals against X is showed below:

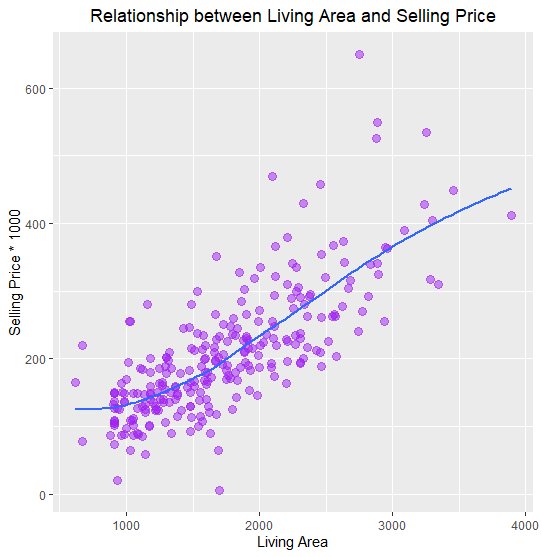
(2) the normal probability plot of semi-studentized residuals against X is showed below:



According to these two plots, we can make sure that, in general, this model fits respectively well because we can tell that most residuals is located between -2 and 2; compared with the majority of the residuals, outliers are not too many.

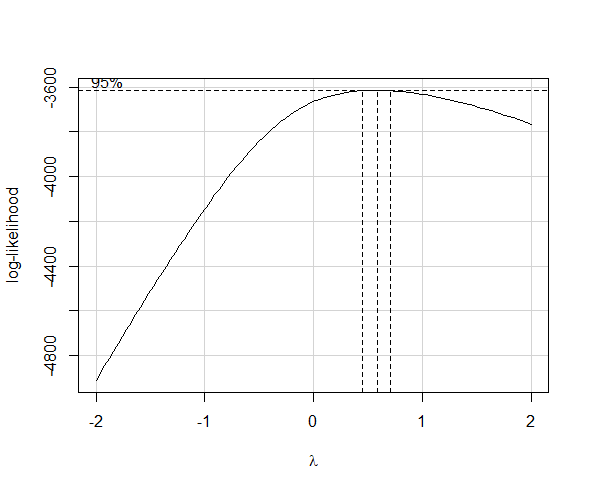
1. Comment on the possible simple transformations of either Y or X, or of both, if any, that can make a linear model more appropriate for the transformed data. Note: there is no need to refit the model to the transformed data.

A: I revised the original straight line into a curved line so that I can observe the breakpoint:

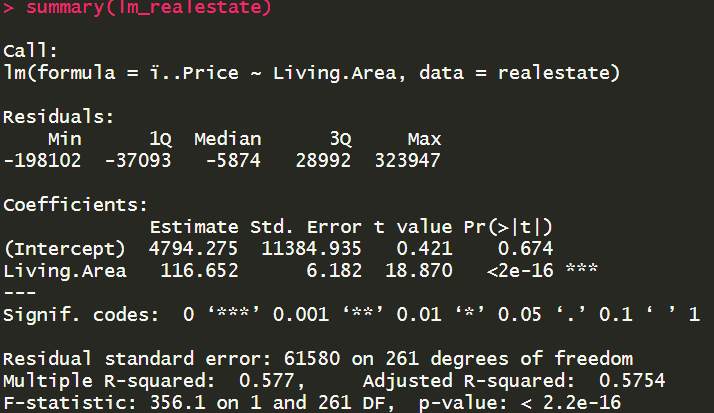


It seems that this curved line is similar to those of some kinds of Logarithmic function.

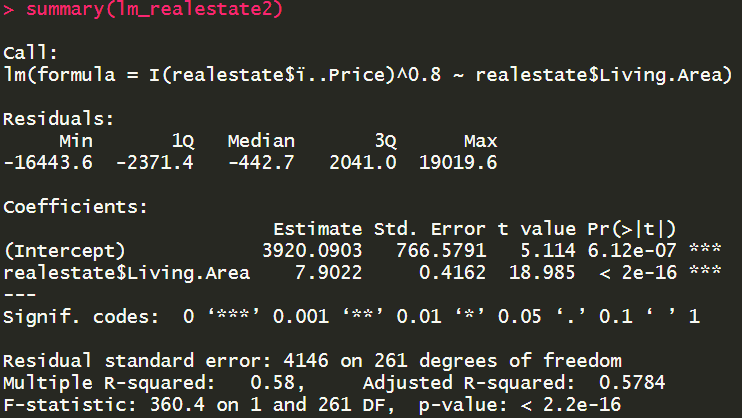
Next, we do some boxcox-transformation tests for the linear model.



According to the result of lambda (close to 0.8), I attempt to choose Y0.8

Here is the original result from summary(lm\_realestate):

Here is the revised result from summary(lm\_realestate2):



Here we can see that although not very much, but we actually make the linear model a little more appropriate than the original version according to the R2 value.