1.

a)

Sample mean and sample variance of both stores can be calculated as follows:

|  |  |  |
| --- | --- | --- |
|  | Miller's | Albert's |
|  | 119.25 | 111.99 |
|  | 121.32 | 114.88 |
|  | 122.34 | 115.11 |
|  | 120.14 | 117.02 |
|  | 122.19 | 116.89 |
|  | 123.71 | 116.62 |
|  | 121.72 | 115.38 |
|  | 122.42 | 114.4 |
|  | 123.63 | 113.91 |
|  | 122.44 | 111.87 |
| Mean | 121.916 | 114.807 |
| Variance | 1.955004 | 3.386712 |

Let = Mean weekly expense of Miller’s = 121.916 and = Mean weekly expense of Albert’s = 114.807

Two samples are assumed to be independent and have equal variances. So, we can calculate the Pooled estimate of Standard Error:

This is a two-tailed hypothesis testing with the mean different between and = 0

Test statistic can be calculated as:

At 0.05 level of significance with degree of freedom = 18, the cutoff t-statistic = 2.101

Since the Test statistic value exceeds the cutoff t-statistic, we reject the null Hypothesis and conclude that the mean weekly expense between Miller’s and Albert’s supermarket chains are different ANS

b)

A 95% confidence interval band can be calculated as follows:

So, at 95% confidence interval, the mean weekly expense different between Miller’s and Albert’s is between $8.645 and $5.573 ANS

c)

Null Hypothesis:

Alternative Hypothesis:

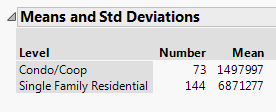
From a), the Test statistic can be calculated as follows:

This is a one-tailed hypothesis testing. At 0.05 level of significance with degree of freedom = 18, the cutoff t-statistic = 1.734

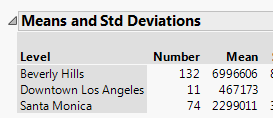
Since the Test statistic value exceeds the cutoff t-statistic, we reject the null Hypothesis and conclude that the mean weekly expense of Miller’s supermarket chain is more than $5 higher than Albert's.ANS

2.

a)



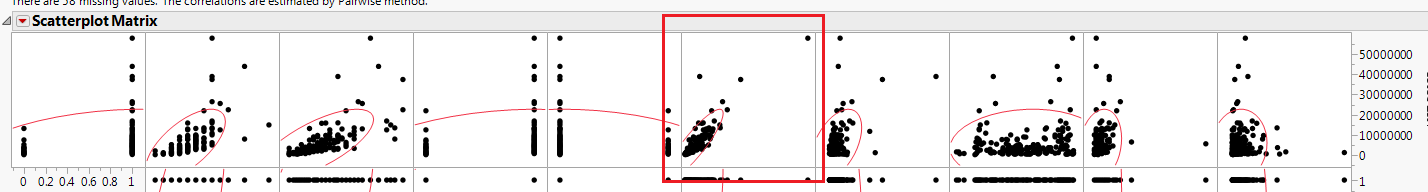
For Home Type, as Condo/Coop estates generally have a lower price, it is set as a baseline. A dummy variable ‘is\_single\_family\_resident’ is created and has a value of 1 if the Home Type is ‘Single Family Residential’, 0 otherwise.



For Location, Downtown LA estates generally have a lower price, 2 dummy variable are crated:

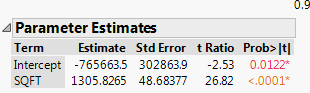
‘is\_Beverly\_Hills’ – equals 1 if the location is Beverly Hills, 0 otherwise.

‘is\_Santa\_Monica’ – equals 1 if the location is Santa Monica, 0 otherwise.



From the correlation matrix, SQFT(Square footage of the living space) seems to be the best predictor. ANS

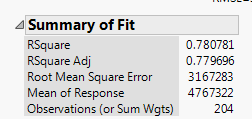
b)



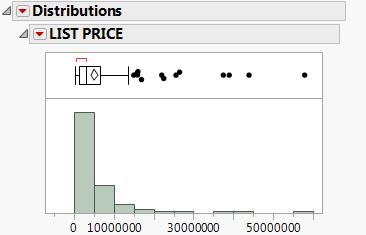
Fit price by SQFT, the regression function is:

ANS

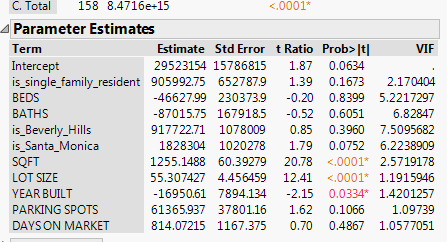
c)



Standard error of the estimate is $3,167,283. This is lower than the standard deviation of Y variable(Price), which is $7,230,747. That is, using SQFT is helpful in predicting the real estate price, but the error is still high because the Price distribution is heavily right skewed:

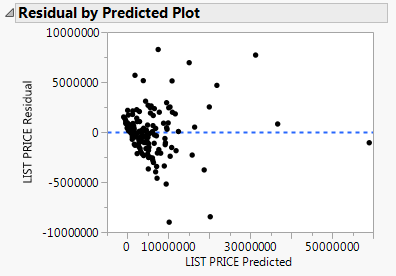
 ANS

d)



By fitting the model to all variables, some variables have Variance Inflation Factor that exceed 5, but none has VIF more than 10. According to the guideline, this shows that multicollinearity exists, but not severe. If we were serious about the multicollinearity, I’d try to drop the variable that has highest VIF first and iteratively repeat the step if VIF of the remained variables are not within the safe area. I will assume it’s a benign problem here. ANS

e)



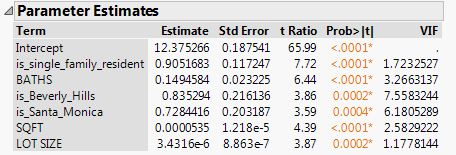
According to the Residual plot, there is a heteroscedasticity issue in the error term. This is likely because of the right skewed distribution of the dependent variable. I’d try to apply a log transformation to the Y(Price) variable. ANS

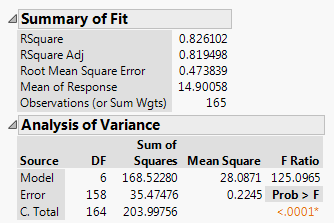
f)

After applying Ln function to the price variable (to adjust for fanned out error term), I then iteratively remove variables that has coefficient with P-value exceed 5% significance level (remove variables that may not play a significant role in the prediction), and carefully inspect the change in the Adjusted R-Square. The process can be summarized as follow:

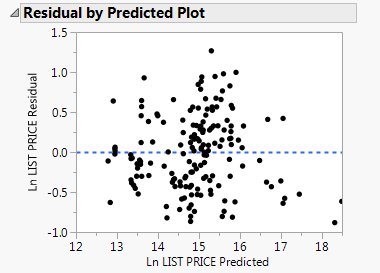
|  |  |  |  |
| --- | --- | --- | --- |
| Removed variable | P-Value | R-Square after removing | Adjusted R-Square after removing |
| None (all variables included) | N/A | 0.83 | 0.82 |
| DAYS ON MARKET | 0.93 | 0.83 | 0.83 |
| YEAR BUILT | 0.73 | 0.83 | 0.82 |
| PARKING SPOTS | 0.12 | 0.83 | 0.82 |
| BEDS | 0.06 | 0.83 | 0.82 |

I get the following model:





The residual plot looks much better:

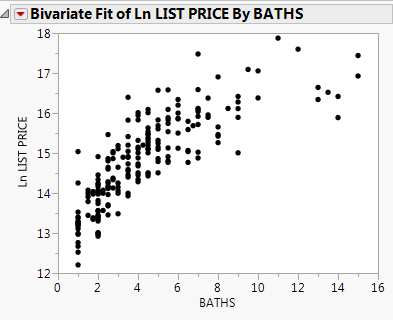


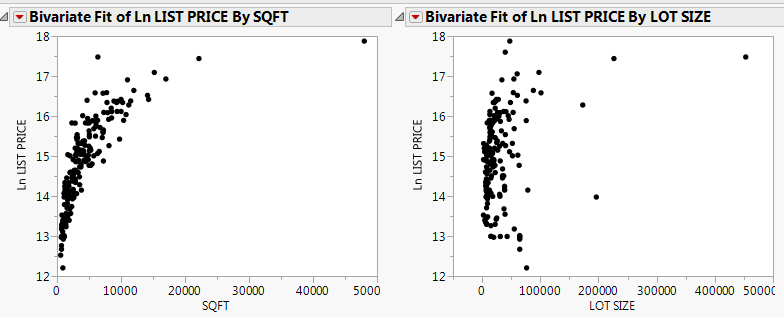
The equation of the final model is:

, where ‘IS SINGLE FAMILY RESIDENT’, ‘IS BEVERELY HILLS’, and ‘IS SANTA MONICA’ are binary variables.

ANS

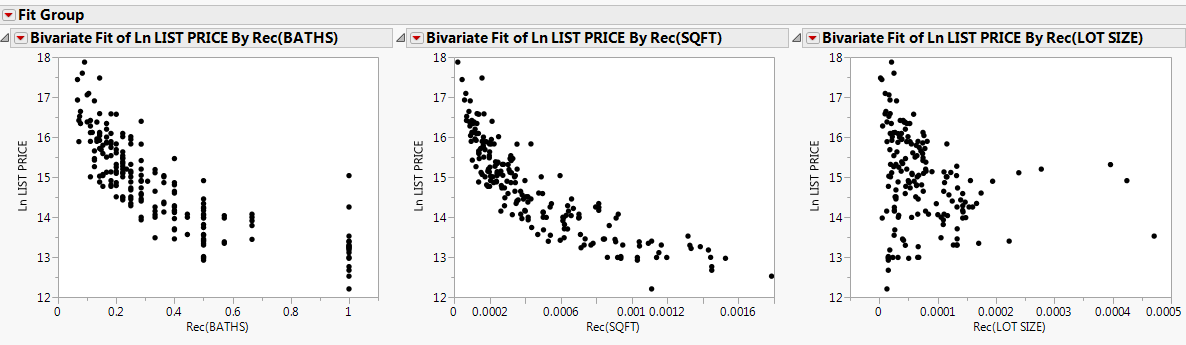
g)



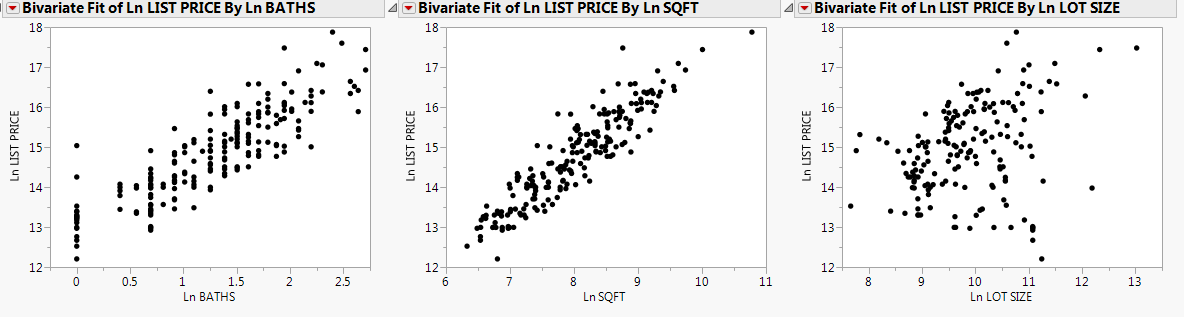


Inspecting the scatter plots, BATHS, SQFT, and LOT SIZE may have a non-linear relationship with Ln(Price), as shown in the above figures. I tried plotting the Ln(Price) variable with their Reciprocal and Log transformation forms to see if any linear relationship can be derived:

Reciprocal Transformation:



Log Transformation:

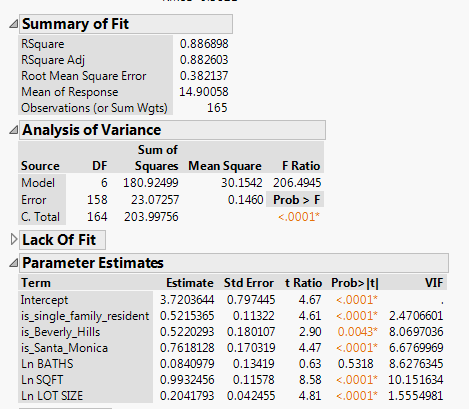


My recommendation is to try applying Log transformation to all 3 variables and refit the model. It does a better job at approximating a linear relationship. Besides, I don’t think we have a theoretical reason to believe that Ln(Price), a % increase in Price, will be bounded by those three variables, which is assumed by the reciprocal model.

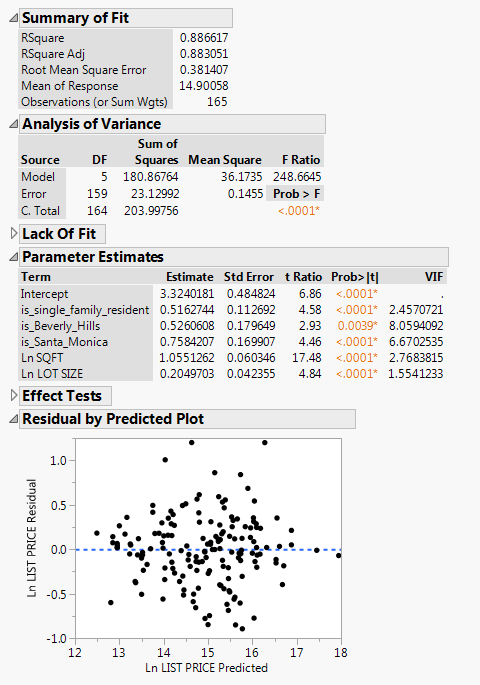
ANS

h)

Fitting the model with the new transformed variable, I get the following data:



As the P-Value of Ln(BATHS) is quite high, I tried dropping it and get the following result (noted: I did not remove Ln(SQFT) even though its VIF exceeds the threshold because it is still quite in a borderline territory and it will be very strange to say that SQFT has no influent on the listed price):



The process can be summarized as follow:

|  |  |  |  |
| --- | --- | --- | --- |
| Removed variable | P-Value | R-Square after removing | Adjusted R-Square after removing |
| None (all variables included) | N/A | 0.89 | 0.88 |
| Ln(BATHS) | 0.53 | 0.89 | 0.88 |

As doing this improved Adjusted R-Square from 0.8826 to 0.8830, Residual plot looks nicer, and VIF of all independent variables looks fine, I decided this to be the final model.

The final model is:

, where ‘IS SINGLE FAMILY RESIDENT’, ‘IS BEVERELY HILLS’, and ‘IS SANTA MONICA’ are binary variables.

ANS

i)

On average, the listed price of Single Family Resident real estate is 52% higher than Condo/ Coop, holding other variables constant.

On average, the listed price of a real estate in Beverly Hills is 53% higher than in Downtown LA, holding other variables constant.

On average, the listed price of a real estate in Santa Monica is 76% higher than in Downtown LA, holding other variables constant.

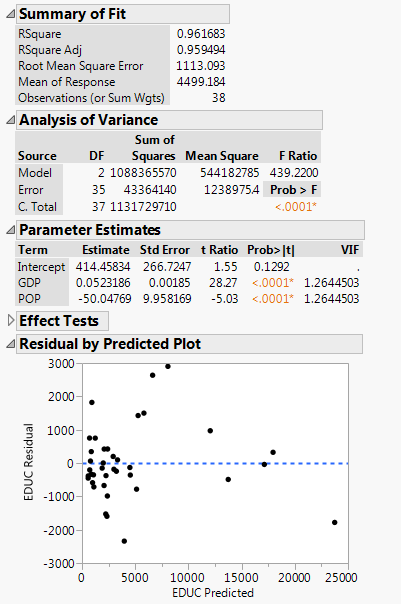
On average, increase a square footage of the living space by 1% resulted in 1.06% increase in the listed price, holding other variables constant.

On average, increase a square footage of the lot property by 1% resulted in 0.20% increase in the listed price, holding other variables constant.

ANS

3.

a) The result of fitting model is as follow:



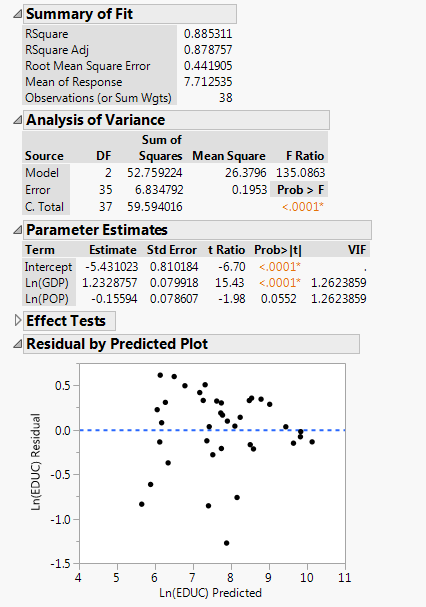
The equation is:

The model has relevant values:

* F-Statistic = 439.22, with P-Value < 0.0001
* t-Statistics of the intercept term = 1.55, with P-Value = 0.13
* t-Statistics of GDP = 28.27, with P-Value < 0.0001
* t-Statistics of POP = -5.03, with P-Value < 0.0001
* R-Square and Adjusted R-Square are around 0.96
* Standard Error of the estimate = 1113

ANS

b) The result of fitting model is as follow:



The equation is:

The model has relevant values:

* F-Statistic = 135.09, with P-Value < 0.0001
* t-Statistics of the intercept term = -6.70, with P-Value < 0.001
* t-Statistics of ln(GDP) = 15.43, with P-Value < 0.0001
* t-Statistics of ln(POP) = -1.98, with P-Value = 0.0552
* R-Square and Adjusted R-Square are around 0.88
* Standard Error of the estimate = 0.44

ANS

c)

* On average, 1% increase in GDP resulted in 1.23% increase in education expenditure, holding other variables constant.
* On average, 1% increase in POP(population) resulted in 0.16% decrease in education expenditure, holding other variables constant.

ANS

d) Even though the R-Square of the log-linear model is lower and the P-value of ln(POP) is a bit on a high side, I think the log-linear is the appropriate model because:

* Residual plot of the log-linear model is more satisfactory. Heteroscedasticity does not look noticeably bad.
* The model’s interpretation sounds more plausible to me; increasing in a number of population has a diminishing effect on a reduction in an education expense. It should not go straight down to zero.

ANS