My first step was to look at scatterplots of the dependent variable as a function of each independent variable individually (excluding the dummy variable representing loans).

|  |  |  |  |
| --- | --- | --- | --- |
| X1 | C:\Users\ellen.bledsoe\OneDrive\Grad_School\Classes\Geography_IntQuantAnalysis\Geog6161\mortpay_vs_income.png | X2 | C:\Users\ellen.bledsoe\OneDrive\Grad_School\Classes\Geography_IntQuantAnalysis\Geog6161\mortpay_vs_sqfoot.png |
| X3 | C:\Users\ellen.bledsoe\OneDrive\Grad_School\Classes\Geography_IntQuantAnalysis\Geog6161\mortpay_vs_type.png | A | C:\Users\ellen.bledsoe\OneDrive\Grad_School\Classes\Geography_IntQuantAnalysis\Geog6161\mortpay_vs_age.png |
| Figure 1: Scatterplots and smoothed best-fit lines for each dependent variable against the independent variable. | | | |

Based on these graphs, I decided to perform transformations on the variables. Initially, I attempted a Box-Cox transformation. Using an optimization function, I found the optimal lambda to be approximately 1.65. When I performed the Box-Cox transformation with this lambda, however, the results were not very different from the original scatterplots. After trying another few transformations, I settled on the natural log of the independent variable. The scatterplots of the transformed Y variable as a function of each dependent variable did not results in fully linear relationships, but the transformation did seem to help somewhat (Figure 2).

|  |  |  |  |
| --- | --- | --- | --- |
| X1 | Macintosh HD:Users:bleds22e:OneDrive:Grad_School:Classes:Geography_IntQuantAnalysis:Geog6161:ln_mortpay_vs_income.png | X2 | Macintosh HD:Users:bleds22e:OneDrive:Grad_School:Classes:Geography_IntQuantAnalysis:Geog6161:ln_mortpay_vs_sqfoot.png |
| X3 | Macintosh HD:Users:bleds22e:OneDrive:Grad_School:Classes:Geography_IntQuantAnalysis:Geog6161:ln_mortpay_vs_type.png | A | Macintosh HD:Users:bleds22e:OneDrive:Grad_School:Classes:Geography_IntQuantAnalysis:Geog6161:ln_mortpay_vs_age.png |
| Figure 2: Scatterplots and smoothed best-fit lines for each dependent variable against the natural log-transformed independent variable. | | | |

My next step was to determine which variables, order of variables and interaction terms I wanted to include in the model. Variables X1 and X2 suggest quadratic or curvilinear relationships. Based on that and pervious studies suggesting that square footage often a second-order polynomial, I decided to include second order variables for X1 and X2 in addition to the original X1 and X2 variables. Looking at the age of housing unit variable (A), it appears to have a cubic relationship. I decided to include both a second and a third order variable of A in addition to the original. Finally, I included numerous interaction terms, mostly focusing on the dummy variable, X3, in order not to miss potentially important interactions. To be on the safe side, I also included a few more possible interaction terms. Therefore, my initial model attempt included the following variables:

Ln(Y) ~ X1, X2, X3, A, X12, X22, A2, A3, X1:X2, X1:X3,

X2:X3, A:X3, A2:X3, A3:X3, A:X1, A:X2, X12:X3, X22:X3

Using all of these potential variables, I ran forward, backwards and bidirectional step-wise ordinary least squares regressions in order to choose a best model; the selection methods used Akaike Information Criteria (AIC) as the selector. I discovered that the forward and bidirectional (starting with the null model) gave me the same best-fit model (AIC = -204.25) while the backwards and bidirectional (starting with the full model) gave me the same best-fit model, but one that was very different from the first (AIC = -203.52). While the AIC values are very similar, the first model had far fewer variables.

I decided to remove some variables to try to get the selection models to show at least some more agreement. This time I decided to be more selective in which variables I used, trying to keep the model more streamlined. Therefore, I selected the original variables, the second-order variables for X1 and X2, the third-order variable for age, and had interaction terms between X3 and all other variables:

Ln(Y) ~ X1, X2, X3, X12, X22, A3, X1:X3, X2:X3, A:X3

This time, after running step-wise regressions forwards, backwards, and bidirectionally both ways, all methods selected the same model, which ended up being the same as the first model selected with the myriad variables included in my very first model (AIC = -204.25). My working model then included:

Ln(Y) ~ X1, X2, X3, X12, X22, A3

Next, I ran an ordinary least squares regression with the new, slimmed down model:

|  |
| --- |
| Residuals:  Min 1Q Median 3Q Max  -0.137267 -0.066210 -0.003324 0.047096 0.230491  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 7.436e+00 6.354e-01 11.703 3.63e-14 \*\*\*  X1 3.619e-02 1.766e-02 2.050 0.047337 \*  X2 -4.134e-03 9.714e-04 -4.256 0.000131 \*\*\*  X1sq -4.486e-04 2.670e-04 -1.680 0.101069  X2sq 2.095e-06 3.625e-07 5.779 1.14e-06 \*\*\*  X3 1.230e-01 3.856e-02 3.189 0.002855 \*\*  Acube -8.919e-05 3.903e-05 -2.285 0.027956 \*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.09628 on 38 degrees of freedom  Multiple R-squared: 0.9527, Adjusted R-squared: 0.9453  F-statistic: 127.6 on 6 and 38 DF, p-value: < 2.2e-16 |
|  |

In my mind, this model was quite good because it didn’t involve an extreme number of variables, most variables show that they have significant effect, and the adjusted R-squared is above 0.9 at 0.945. In this case, all of the variables come out as being significant except for X1. This indicated to me that I did not need to keep that variable in the model, so I reran it.

|  |
| --- |
| Residuals:  Min 1Q Median 3Q Max  -0.179825 -0.069034 -0.003544 0.069564 0.233390  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 7.656e+00 6.362e-01 12.033 1.06e-14 \*\*\*  X1 6.901e-03 2.885e-03 2.392 0.021678 \*  X2 -3.626e-03 9.446e-04 -3.839 0.000442 \*\*\*  X2sq 1.835e-06 3.355e-07 5.470 2.82e-06 \*\*\*  X3 1.220e-01 3.945e-02 3.093 0.003654 \*\*  Acube -9.260e-05 3.988e-05 -2.322 0.025525 \*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.0985 on 39 degrees of freedom  Multiple R-squared: 0.9492, Adjusted R-squared: 0.9427  F-statistic: 145.8 on 5 and 39 DF, p-value: < 2.2e-16 |

* Natural log transformation on Y variable
* Decide which variables and interacts I want to use
* Use both step-wise and AIC to determine best model
  + After multiple iterations, the models from all directions agree on:
  + X1, X2, X3, X1sq, X2sq, Acube
* Check for collinearity
* Test assumptions