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* |

Removing the one variable (X1) dropped by adjusted R-squared value by less than 0.003, which is a miniscule amount. The X1 variable did not seem to be carrying much weight, so I felt justified in removing it from the model.

|  |
| --- |
| Macintosh HD:Users:bleds22e:OneDrive:Grad_School:Classes:Geography_IntQuantAnalysis:Geog6161:pred_vs_resid_mod2.pngMacintosh HD:Users:bleds22e:OneDrive:Grad_School:Classes:Geography_IntQuantAnalysis:Geog6161:qq_mod2.png |
| *Figure 3. Plots of the Y values predicted by the model against the residuals and a QQ plot.* |

To test the assumptions of the model, I first produced a plot of the residuals of the model against the predicted Y values (Fig. 3). I also make a QQ plot (Fig. 3). While the residuals vs. predicted values plot looks fairly normal (the QQ plot seems borderline but probably okay), these plots are only visually diagnostic tests of the assumptions. In addition to these plots, I also ran a few statistical tests of the assumptions:

*Normality:*

I ran a Shapiro-Wilk test for normality on the residuals of the model. The test returned a W = 0.9719 and a p-value = 0.3388. Because the Shapiro-Wilk tests for a null hypothesis that the data are normal, the given p-value indicates that we fail to reject normality, meaning that the residuals are normally distributed.

*Homoscedasticity:*

I ran a Bruesch-Pagan test for homogeneity of variances. This returned values of BP = 5.17 and p-value = 0.3954. Because the null for Breusch-Pagan test is that error variances are homogenous and my p-value was greater than 0.05, I fail to reject the null, suggesting that my variances are homogenous.

Finally, I decided to look and see if any of the points had excessive leverage on the dataset. The DFFITS, Cook’s distance and hat values are listed below (I apologize, I can’t seem to figure out a way to turn this output into a table or dataframe that would make it much easier to read):

dfb.1\_ dfb.X1 dfb.X2 dfb.X2sq dfb.X3 dfb.Acub dffit cov.r cook.d hat inf

1 0.21459 -0.11684 -0.181157 0.17735 0.026128 -2.26e-01 0.35773 1.263 2.15e-02 0.1645

2 -0.37899 0.01243 0.346982 -0.32957 -0.042307 2.84e-01 -0.47855 1.351 3.84e-02 0.2318

3 0.01545 0.13172 -0.036254 0.03079 0.016223 -2.54e-02 -0.25516 1.026 1.08e-02 0.0523

4 0.00789 -0.00249 -0.006408 0.00582 -0.000624 -1.07e-02 0.01673 1.280 4.79e-05 0.0871

5 0.23651 0.15743 -0.210553 0.17030 -0.063871 -3.01e-01 0.50982 0.843 4.15e-02 0.0866

6 0.24633 0.79444 -0.322882 0.25368 0.032664 1.85e-01 -0.87250 1.384 1.25e-01 0.3331

7 -0.01033 -0.03595 0.014006 -0.01109 -0.000853 -2.51e-03 0.04093 1.396 2.86e-04 0.1638

8 0.04954 -0.20751 -0.055112 0.10113 -0.402374 2.45e-02 -0.57435 1.017 5.37e-02 0.1460

9 0.02568 0.00119 -0.033362 0.04181 -0.138788 3.09e-02 -0.17797 1.294 5.39e-03 0.1271

10 -0.05527 0.09570 0.061636 -0.07964 -0.107143 -1.16e-01 0.26631 1.065 1.18e-02 0.0648

11 0.01009 0.04158 -0.017433 0.01604 0.012123 3.72e-02 -0.07447 1.257 9.47e-04 0.0798

12 0.11190 -0.05940 -0.116177 0.12652 0.125614 8.74e-02 -0.26181 1.073 1.14e-02 0.0653

13 0.00767 -0.01985 -0.008342 0.01283 -0.040312 7.33e-03 -0.05970 1.352 6.09e-04 0.1384

14 0.12226 0.14497 -0.120099 0.09699 -0.023043 -4.23e-03 0.26515 1.117 1.18e-02 0.0791

15 0.00965 -0.05751 0.006411 -0.00888 -0.023533 2.40e-02 0.23305 1.005 8.99e-03 0.0417

16 0.06736 0.03007 -0.074747 0.07846 -0.178015 1.24e-02 -0.23737 1.223 9.52e-03 0.1103

17 0.35011 0.12399 -0.345921 0.32581 0.062863 1.43e-01 0.52497 1.146 4.55e-02 0.1712

18 0.06211 0.04019 -0.064847 0.06056 0.016300 1.32e-01 0.19841 1.384 6.70e-03 0.1790

19 -0.01227 0.00107 0.011503 -0.01120 -0.002576 1.40e-04 -0.01624 1.376 4.51e-05 0.1507

20 -0.05775 0.03437 0.048043 -0.04656 -0.008218 -5.33e-02 -0.18652 1.150 5.87e-03 0.0630

21 0.00580 0.02264 -0.010498 0.01024 0.008912 1.65e-02 -0.05024 1.223 4.31e-04 0.0513

22 0.05374 -0.17011 -0.046116 0.06679 0.081964 -7.12e-02 -0.26054 1.100 1.14e-02 0.0723

23 0.10031 0.04401 -0.109245 0.10762 0.078939 5.06e-02 -0.18884 1.143 6.01e-03 0.0614

24 0.07126 -0.00849 -0.071056 0.07037 0.072317 2.30e-02 -0.12780 1.218 2.78e-03 0.0721

25 0.04139 -0.03894 -0.038584 0.04175 0.044548 5.03e-04 -0.08245 1.262 1.16e-03 0.0846

26 -0.00998 0.01850 0.008946 -0.01089 -0.012933 -1.48e-03 0.02776 1.305 1.32e-04 0.1054

27 0.23479 -0.15117 -0.222996 0.23303 0.237489 6.85e-04 -0.42531 0.906 2.93e-02 0.0767

28 -0.04327 -0.05621 0.044391 -0.03699 0.003954 8.51e-02 0.10919 2.201 2.04e-03 0.4701 \*

29 0.01351 -0.09744 -0.010372 0.02269 0.046384 2.85e-02 -0.12767 1.310 2.78e-03 0.1238

30 0.09375 -0.12343 -0.073287 0.07925 0.024490 -3.52e-01 -0.38712 1.728 2.54e-02 0.3523 \*

31 -0.06274 -0.11275 0.064675 -0.04885 0.018186 -6.93e-02 -0.21515 1.165 7.81e-03 0.0788

32 0.07947 -0.03755 -0.072676 0.06977 0.107790 -6.95e-05 -0.17008 1.207 4.91e-03 0.0805

33 -0.10954 0.02304 0.105151 -0.10149 -0.127499 -1.76e-02 0.20865 1.163 7.34e-03 0.0755

34 -0.41780 -0.02895 0.421039 -0.40876 -0.392217 -1.06e-01 0.71094 0.480 7.36e-02 0.0710 \*

35 -0.05709 0.02020 0.057265 -0.06242 0.154994 1.08e-02 0.22003 1.216 8.19e-03 0.1015

36 0.00203 0.00151 -0.002137 0.00201 -0.007268 3.19e-05 -0.01018 1.293 1.77e-05 0.0965

37 -0.05189 -0.06989 0.056137 -0.04881 0.188109 5.14e-03 0.26470 1.178 1.18e-02 0.1008

38 -0.00411 0.00988 0.000299 0.00343 -0.042668 -2.19e-03 0.06823 1.298 7.95e-04 0.1056

39 -0.09088 0.15491 0.047519 -0.02111 -0.352169 1.31e-01 0.58040 0.875 5.38e-02 0.1118

40 0.01570 -0.01550 -0.017717 0.02282 0.041055 1.79e-02 0.07600 1.358 9.87e-04 0.1443

41 -0.34518 0.05562 0.367655 -0.40970 -0.326340 -6.02e-03 -0.73989 0.983 8.80e-02 0.1826

42 -0.00186 -0.00039 0.002069 -0.00226 -0.002592 -2.62e-04 -0.00526 1.357 4.73e-06 0.1386

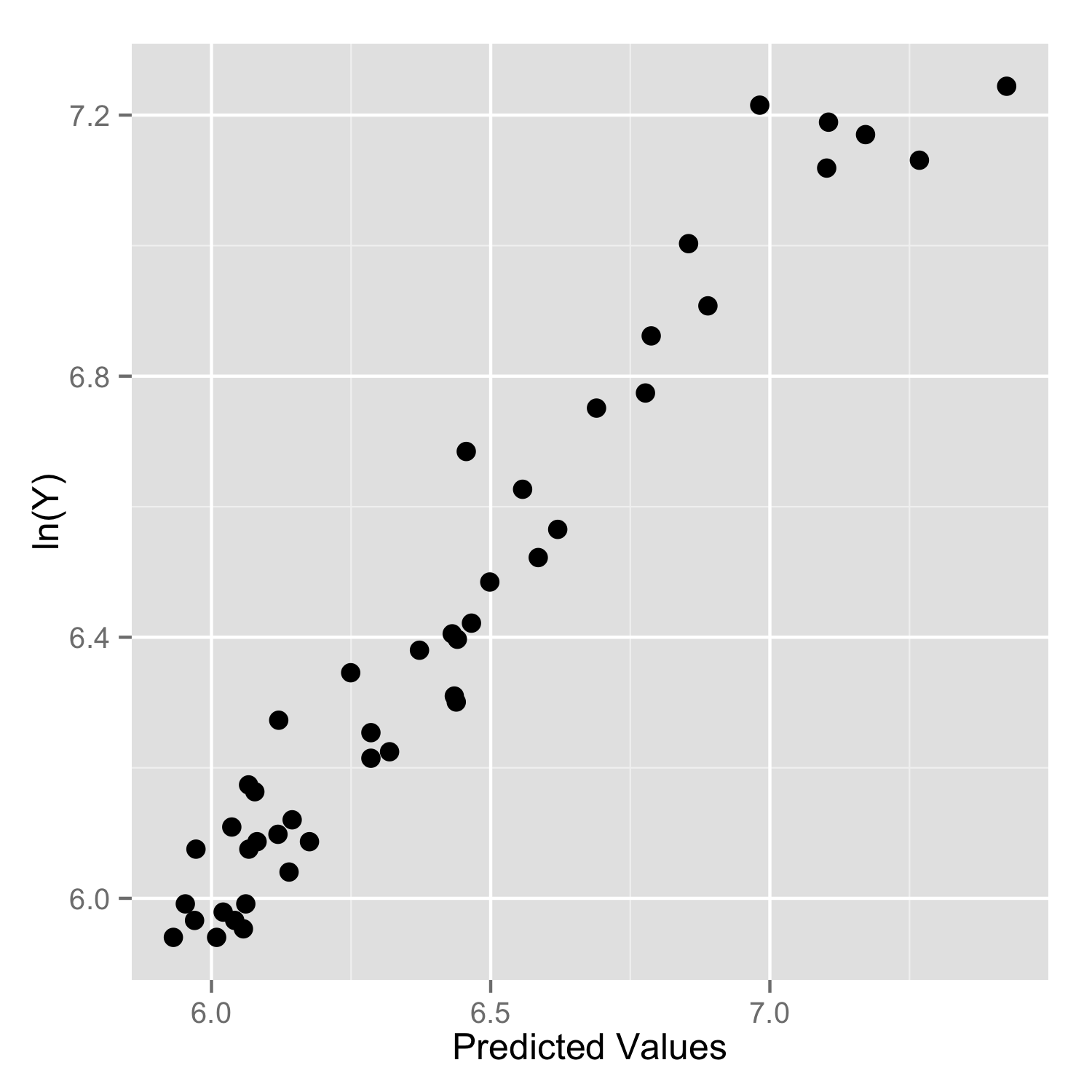
43 0.06726 -0.09828 -0.079736 0.11346 0.573725 2.48e-02 0.90321 0.456 1.17e-01 0.1012 \*

44 -1.12428 -0.26228 1.236796 -1.35634 0.716365 1.11e-01 -2.00458 0.775 5.91e-01 0.3925 \*

45 0.12941 -0.15998 -0.141051 0.19359 -0.201259 -3.25e-02 0.50069 1.285 4.19e-02 0.2131

A couple of the values look a bit high, such as the hat value for observations 28 and 44. The other values, however, seem to be fairly consistent, so I’m going to assume that they are not a big enough problem to warrant taking them out.

After all of this, I feel pretty confident in my model. The actual values of ln(Y) seem to match up fairly well with the predicted values of ln(Y), as shown below.



*Recovered Estimate and Mean Predicted Interval Band for Value #25:*

I can figure out that the predicted value of value 25: lnY predicted = 6.431090. In transforming backwards, we get that the predicted Y for value 25 is $620.85, as compared to an original value of $605 for monthly mortgage payments. The confidence band is lower = 6.373123 and upper = 6.489056. Back transformed, they equal $585.88 and $657.9.