After estimating various equations in our attempt to find the determinants of the log sale prices using TOTALBSMTSF and BEDROOMABVGR series as explanatory variables, we arrived at the following model:



Trying to explain the log sale prices using log of the total basement footage prompted us to restrict our sample only to observations for which positive footage was given. Thus, we removed the 37 houses with a reported basement of 0 square feet from our sample to allow for the log estimation comparison with the model using the level of footage (note that we still have a large sample left). Further improvement was achieved by expanding the bedroom series into several dummies with zero bedrooms in the basement as the benchmark.

The RESET test rejected the null of linearity, therefore we added the interaction term as well as the square of the basement footage to remedy for misspecification.



The Jarque-Bera statistic led us to rejecting the hypothesis of normally distributed residuals and the White test rejected their homoskedasticity. The equation was therefore augmented with heteroskedasticity consistent standard errors.

Using the restricted sample, the model without the log transformation slightly outperformed the log one due to its relatively higher R-squared and lower information criteria.

To summarize, our model explains approximately 47 per cent of the log sale price variation—the F test results in the overall significance of the explanatory variables. Except for the dummy indicating one bedroom at the basement level, all variables are also individually significant at the 5% level. The estimated effect of total basement footage is evaluated to have a positive impact on log sale prices, but it is decreasing in both the total basement footage and in the number of bedrooms at the basement level. Expanding the series bedroomabvgr led us to estimate that houses with the same basement footage are expected to cost more with each additional bedroom at the basement level, with the only exception being that the effect of bedroomabvgr=5 is estimated to be greater than bedroomabvgr=6. Overall, the results are in line with our intuition.

The big data selection methods allowed us to produce the following model:



Even though the model includes too many variables for it to be easily justified by economic theory, we chose this one as it had the highest adjusted R squared and lowest information criteria compared to other selection methods.

To elaborate on our selection process, we considered all the EViews VARSEL options: uni-directional, stepwise, swapwise, combinatorial, GETS and Lasso. The output of the lastly mentioned method is given below for illustration.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dependent Variable: LOG(SALEPRICE) | | | |  |
| Method: Variable Selection | | |  |  |
| Date: 02/25/23 Time: 15:33 | | | |  |
| Sample: 1 1460 | | |  |  |
| Included observations: 1423 | | | |  |
| Number of always included regressors: 41 | | | | |
| No search regressors | | |  |  |
| Selection method: Lasso | | |  |  |
| Lambda at minimum AIC: NA | | |  |  |
| Huber-White-Hinkley (HC1) heteroskedasticity consistent standard | | | | |
| errors and covariance | | | |  |
|  |  |  |  |  |
|  |  |  |  |  |
| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|  |  |  |  |  |
|  |  |  |  |  |
| C | 4.850735 | 0.740982 | 6.546361 | 0.0000 |
| TOTALBSMTSF^2 | -1.03E-07 | 4.65E-09 | -22.06037 | 0.0000 |
| BEDROOMABVGR=1 | -0.040210 | 0.023352 | -1.721890 | 0.0853 |
| BEDROOMABVGR=3 | 0.053458 | 0.013942 | 3.834365 | 0.0001 |
| BEDROOMABVGR=4 | 0.086241 | 0.028451 | 3.031195 | 0.0025 |
| BEDROOMABVGR=5 | 0.047465 | 0.051331 | 0.924691 | 0.3553 |
| BEDROOMABVGR=6 | 0.107856 | 0.074947 | 1.439098 | 0.1503 |
| TOTALBSMTSF\*BEDROOMABVGR | -4.43E-05 | 9.72E-06 | -4.554507 | 0.0000 |
| LOTAREA | 2.63E-06 | 4.84E-07 | 5.434233 | 0.0000 |
| OVERALLQUAL | 0.060714 | 0.004891 | 12.41324 | 0.0000 |
| OVERALLCOND | 0.063222 | 0.003917 | 16.13904 | 0.0000 |
| YEARBUILT | 0.002850 | 0.000364 | 7.827759 | 0.0000 |
| TOTALBSMTSF | 0.000540 | 3.51E-05 | 15.36765 | 0.0000 |
| GRLIVAREA | 0.000284 | 2.26E-05 | 12.55591 | 0.0000 |
| GARAGECARS | 0.056512 | 0.007818 | 7.228093 | 0.0000 |
| NEIGHBORHOOD="Blueste" | -0.097278 | 0.020248 | -4.804317 | 0.0000 |
| NEIGHBORHOOD="BrDale" | -0.150588 | 0.028197 | -5.340647 | 0.0000 |
| NEIGHBORHOOD="BrkSide" | 0.008454 | 0.033253 | 0.254239 | 0.7993 |
| NEIGHBORHOOD="ClearCr" | 0.074787 | 0.030983 | 2.413766 | 0.0159 |
| NEIGHBORHOOD="CollgCr" | 0.029998 | 0.014264 | 2.103044 | 0.0356 |
| NEIGHBORHOOD="Crawfor" | 0.130214 | 0.028531 | 4.563943 | 0.0000 |
| NEIGHBORHOOD="Edwards" | -0.056609 | 0.035923 | -1.575847 | 0.1153 |
| NEIGHBORHOOD="Gilbert" | 0.033877 | 0.016536 | 2.048725 | 0.0407 |
| NEIGHBORHOOD="IDOTRR" | -0.128856 | 0.050955 | -2.528812 | 0.0116 |
| NEIGHBORHOOD="MeadowV" | -0.140536 | 0.029902 | -4.699872 | 0.0000 |
| NEIGHBORHOOD="Mitchel" | -0.025551 | 0.023753 | -1.075704 | 0.2822 |
| NEIGHBORHOOD="NAmes" | -0.000172 | 0.021521 | -0.007990 | 0.9936 |
| NEIGHBORHOOD="NoRidge" | 0.109734 | 0.025695 | 4.270669 | 0.0000 |
| NEIGHBORHOOD="NPkVill" | -0.064604 | 0.021731 | -2.972855 | 0.0030 |
| NEIGHBORHOOD="NridgHt" | 0.115876 | 0.024712 | 4.689049 | 0.0000 |
| NEIGHBORHOOD="NWAmes" | -0.026447 | 0.020268 | -1.304883 | 0.1921 |
| NEIGHBORHOOD="OldTown" | -0.090898 | 0.031162 | -2.916962 | 0.0036 |
| NEIGHBORHOOD="Sawyer" | -0.007727 | 0.025525 | -0.302720 | 0.7621 |
| NEIGHBORHOOD="Somerst" | 0.069680 | 0.018555 | 3.755221 | 0.0002 |
| NEIGHBORHOOD="StoneBr" | 0.144373 | 0.034405 | 4.196278 | 0.0000 |
| NEIGHBORHOOD="SWISU" | 0.010096 | 0.035183 | 0.286964 | 0.7742 |
| NEIGHBORHOOD="Timber" | 0.039547 | 0.022467 | 1.760199 | 0.0786 |
| NEIGHBORHOOD="Veenker" | 0.091703 | 0.039467 | 2.323530 | 0.0203 |
| BSMTQUAL="Fa" | -0.083714 | 0.035300 | -2.371487 | 0.0179 |
| BSMTQUAL="Gd" | -0.084019 | 0.017443 | -4.816865 | 0.0000 |
| BSMTQUAL="TA" | -0.093802 | 0.020799 | -4.509958 | 0.0000 |
|  |  |  |  |  |
|  |  |  |  |  |
| R-squared | 0.893986 | Mean dependent var | | 12.03691 |
| Adjusted R-squared | 0.890918 | S.D. dependent var | | 0.393816 |
| S.E. of regression | 0.130068 | Akaike info criterion | | -1.213129 |
| Sum squared resid | 23.38025 | Schwarz criterion | | -1.061561 |
| Log likelihood | 904.1414 | Hannan-Quinn criter. | | -1.156516 |
| F-statistic | 291.3512 | Durbin-Watson stat | | 1.996329 |
| Prob(F-statistic) | 0.000000 |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  | |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

Estimating the equation of both the small OLS model and our preferred big data selection method model on a restricted sample of 1400 observations and forecasting the remaining 60 produced the following results:



The RMSE of the smaller model was 52305.45$, while the bigger model had RMSE of 19356.74$. When comparing the two, we need to take note of the difference in included observations, which was 60 for the smaller model and 55 for the bigger one due to unavailability of data. However, it is still likely that the big data methods specified model performs better in terms of forecasting as the magnitude of its RMSE is less than half of the small OLS model.

Finally, using the given house characteristics,

* LotArea: 8500
* OverallQual: 7
* OverallCond: 5
* YearBuilt: 2003
* TotalBsmtSF: 1000
* GrLivArea: 1700
* FullBath: 2
* BedroomAbvGr: 3
* GarageCars: 2

the big data model produced a

* point prediction of 127 541.9$,
* with 95% CI in the range from 94 100$ to 173 000$,

where the rest of the necessary explanatory variables were added to the observation 1461 based on the most frequent characteristics of the rest of the sample.

For comparison, the small OLS model produced a point prediction 165 648.3$ and 95% CI from 93 500$ to 294 000$, where the width of the interval illustrates that the bigger model produces relatively more certain forecasts.

The boxplot of SalePrice below shows that the forecast is near the mean and median price of the houses sold in our data.

