Part-II-Writeup

Group 5

Part 1: Introduction

For this project, we are being asked to explain which factors played an important role on a selection of paintings sold in Paris in the late 18th century. We will use these factors to identify undervalued/overvalued paintings in the dataset. To facilitate this, we have been provided with auction price data from 1764-1780 containing information on painting/artist characteristics and the sale itself. First, we will use EDA to get to know the data. This will help us get an initial sense of which variables are associated with price, and which variables to include in our preliminary models.

After EDA is complete, we will move on to the model building portion of the analysis. Using the important variables identified in EDA, we will create a multiple linear regression model attempting to explain price as a function of these covariates. After considering potential interactions between the selectors, we will try to further improve/simplify our model via variable selection with AIC/BIC. Finally, we will validate our model by checking the assumptions of linear regression, and assert that we have not made any major violations. How well our model performs will be a result of several factors:

- Does our model correctly capture what the true connection between variables and price is? I.e., does our model have low bias?
- How well does our model fit the data relative to the null model? A common metric for this is the root mean squared error (RMSE), which is a function of the residuals of our model. We are looking for a low RMSE on “new” data (the test dataset).
- Does our model perform well on new data? I.e., does our model have low variance? If we add too many predictors, we risk overfitting our model to the training data and jeopardizing performance on new data.
- Does our model “cover” the predictions well? From our model, we can create prediction intervals for where we expect the price to fall. We would like the true value to be covered most, but not all of the time. If we’re covering the true value all of the time, our prediction intervals are too wide, and our prediction intervals are not as precise as we would like.

After our analysis with linear regression, we will proceed forward with a more complicated model. This will be contained in the subsequent writeup.

For the second part of this project, we are focusing on obtaining predictive accuracy as the primary objective. We are allowed to lose some interpretation if the prediction is better. We will investigate splines, BMA, etc. as these are powerful tools that can still be somewhat interpreted.

Part 2: EDA

To begin with the EDA, first we got a summary of the variables and discard any unneeded portions. Our response variable is `logprice`. As such, we will not use `price` in our analysis. `count` is a column of all 1’s, and provides no value. The data is a mix of continuous, binary and categorical variables. The variables `position`, `nfigures` and the ones that are related to the dimensions of the painting(`Surface`, `Width_in` etc) are all continuous. `year` can be taken either as continuous or as factors, we choose to use it as a factor. The variable `lot` is a number so can be used as numeric. The variable pertaining to features of the paintings(like `lrgfont`, `engraved` etc), `diff_origin`, `artist_living` and `Interm` are binary and we will use them as factors. All the other variables are multilevel categorical variables.

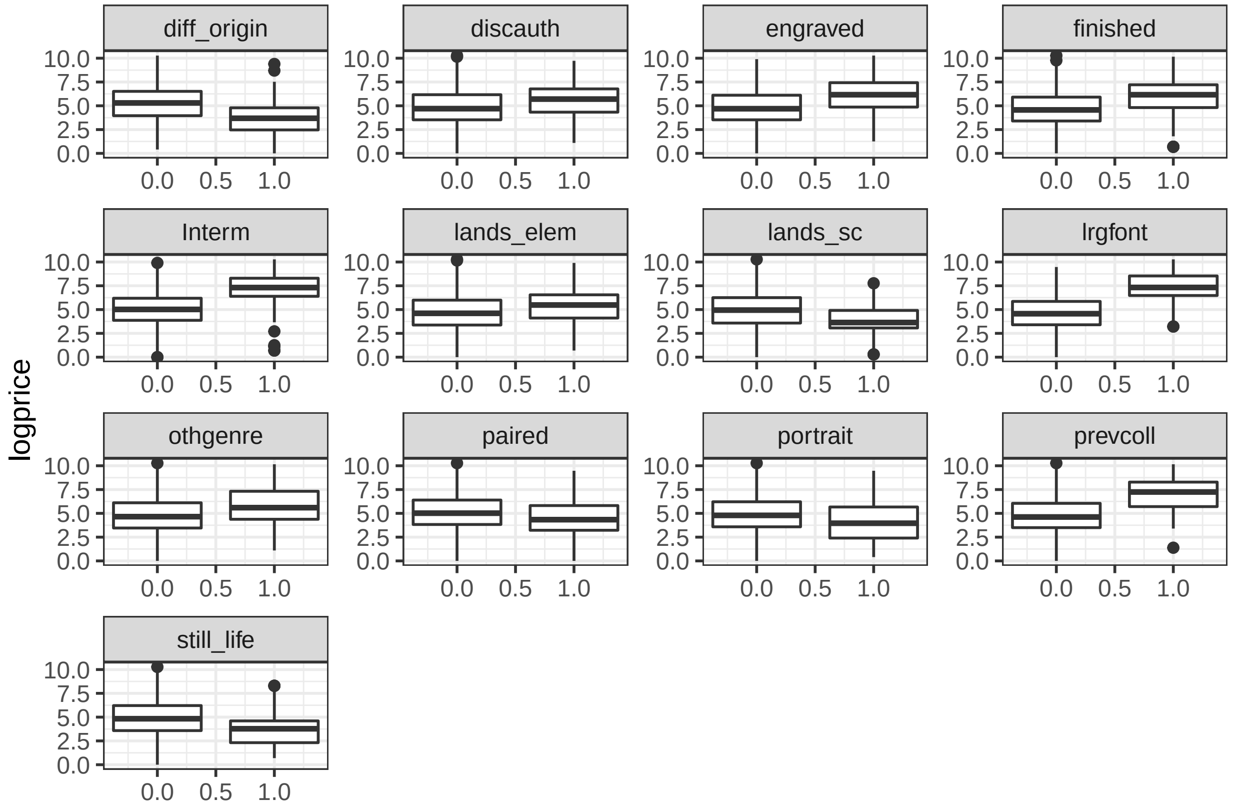


Figure 1: Boxplots between logprice and some binary variables

Investigating Binary Variables

Next, we note that a large amount of the variables are binary. We can investigate the association with price via boxplots:

Figure 1 provide a great visualization on how a given binary variable is associated with changes in logprice. It's hard to quantify how much of a difference in price should warrant variable inclusion, but as we will be performing variable selection later, it shouldn't hurt to include more than necessary. We see the following potential predictors from the graphs:

- `prevcoll`
- `lrgfont`
- `diff_origin`
- `finished`
- `engraved`
- `lands_sc`
- `portrait`
- `paired`
- `still_life`
- `lands_elem`
- `Interm`
- `figures`
- `discauth`
- `othgenre`

We note that many of these binary variables seem to be strongly associated with `logprice`, so we will consider them further when we are building the initial model.

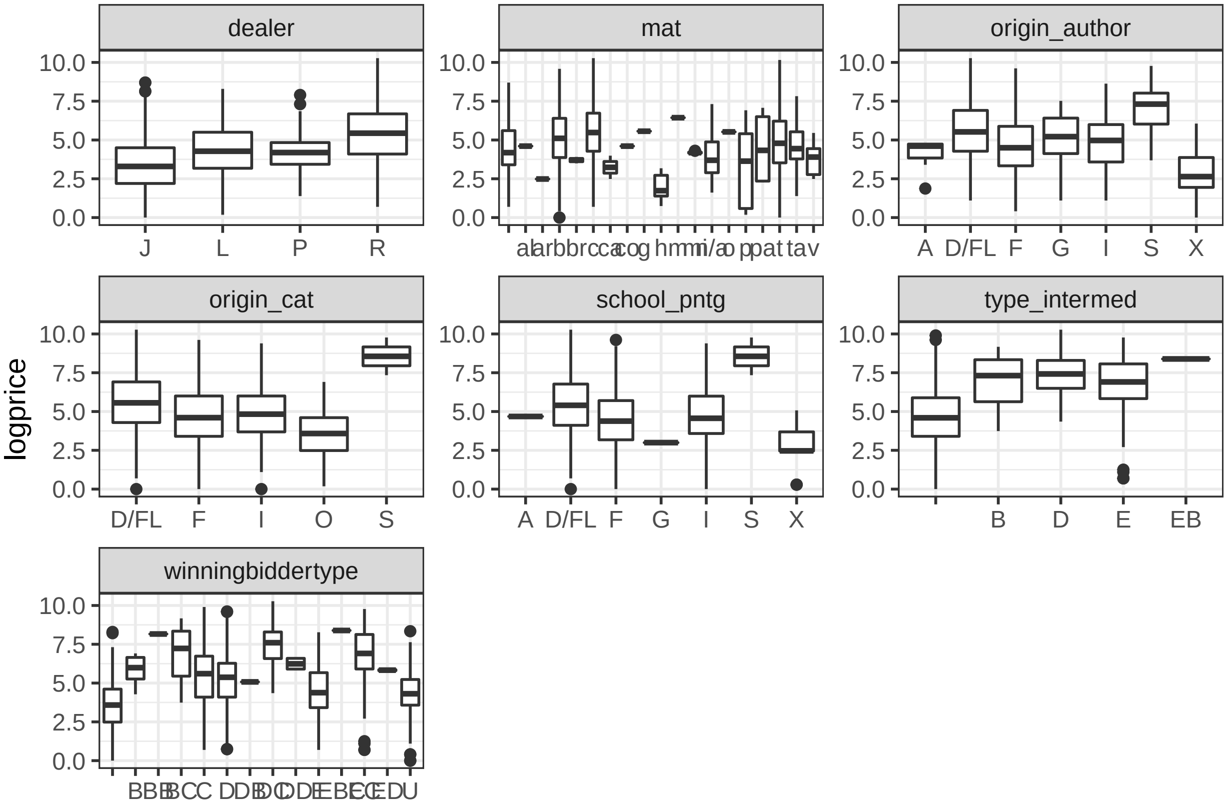


Figure 2: Boxplots between logprice and some categorical variables

The variable `Interm` has many missing values corresponding to the winning bidder being "Unknown", and we found that including that "missingness" as a level, the logprice was lower than the when we know whether there is an intermediary or not. Surprisingly, for the same observations, `type_intermed` is just blank, same as when `Interm` is 0, not missing. So we can just use `type_intermed` instead of `Interm` and assume that the blanks in place of NAs are imputations, or maybe that's how it was supposed to be coded.

We also note that the variable `peasant` and `othgenre` are like two levels of a factor(they cannot be 1 together as described in the codebook, and both describe the scene of the painting). So, while we don't think `peasant` has any significant impact on `logprice`, we may include it anyway.

Investigating categorical variables

Several variables have multiple categories (country origin, type of endbuyer, etc.). We can visualize these with box plots, there will just be more boxes than a binary variable.

Several of the categories for variables have very small sample sizes, and as such cannot be used for our analysis (`mat`, `school_pntg`, `winningbiddertype`, `type_intermed`). The other variables initially seem feasible. It would make sense that nationality could be an important factor; perhaps certain countries are known more for their art. Similarly, it makes sense that `year` could be associated with price. Perhaps based on the economic circumstances, people would be more willing to spend different amounts of money on art. There doesn't seem to be a linear trend, however, so we will need to treat `year` as a categorical variable rather than numeric.

We will also choose only `origin_author` from `origin_author`, `origin_cat` and `school_pntg`, because they are likely to be the same in many observations and out of the three, `origin_author` has all the categories and has reasonable amount of observations for each category compared to `school_pntg`.

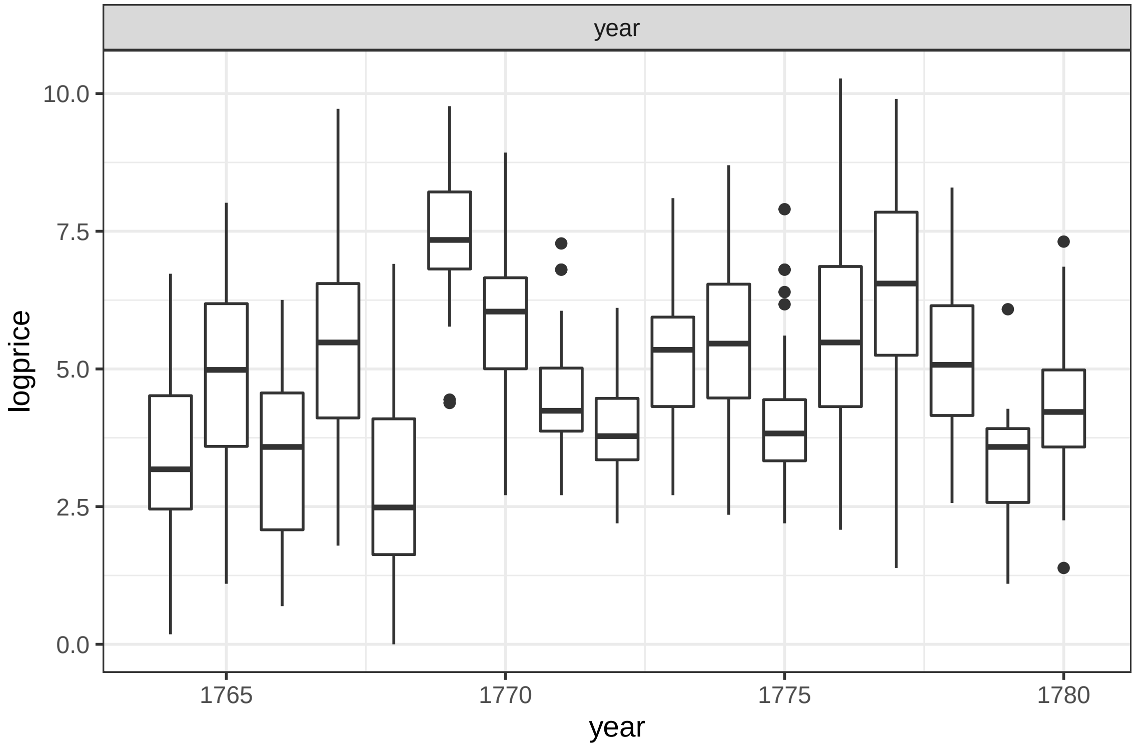


Figure 3: Boxplots between logprice and year

Continuous Variables

Finally, certain variables are continuous in nature. We can investigate the relationship with price via scatterplots.

- Note: For the `position` variable, there were a few observations whose values were not between 0 and 1. We assumed this was an error caused by percentage being entered instead of fractions. These values were corrected by dividing by 100.

In no way does `position` seem to be associated with price, so we will not consider it further. We expected `Surface` to play a much larger role in price. Bigger paintings have more expensive materials and take artists longer to finish. We see from the graph, however, that there is only a very slight increase in `logprice` as `Surface` increases. As such, we will not include for now as it doesn't seem to be as related as the other variables. Additionally, the test data contains several NA values for `Surface` compared to the other variables, and improper imputation could be misleading.

EDA between Predictors

Apart from the EDA for the predictors vs. `logprice`, on looking at the relationships between the predictors, we found a few things which seem important:

- Only dealer "R" can have `lrgfont` = 1. It also looks like the dealer is most likely "R" when `engraved` or `prevcoll` are 1.
- The variable `lands_sc` and `lands_elem` are never 1 together in the training data, but there very few cases in the test data. We observe the same in the variables `portrait` and `still_life`.
- There is only one observation with `type_intermed` "EB" along with `winningbiddertype` being "EBC", same is the case in test data. Since, there is only one observation of "EB", it will be difficult to use `type_intermed` unless we change something.
- Whenever there is an intermediary involved, the end buyer is most likely a collector.

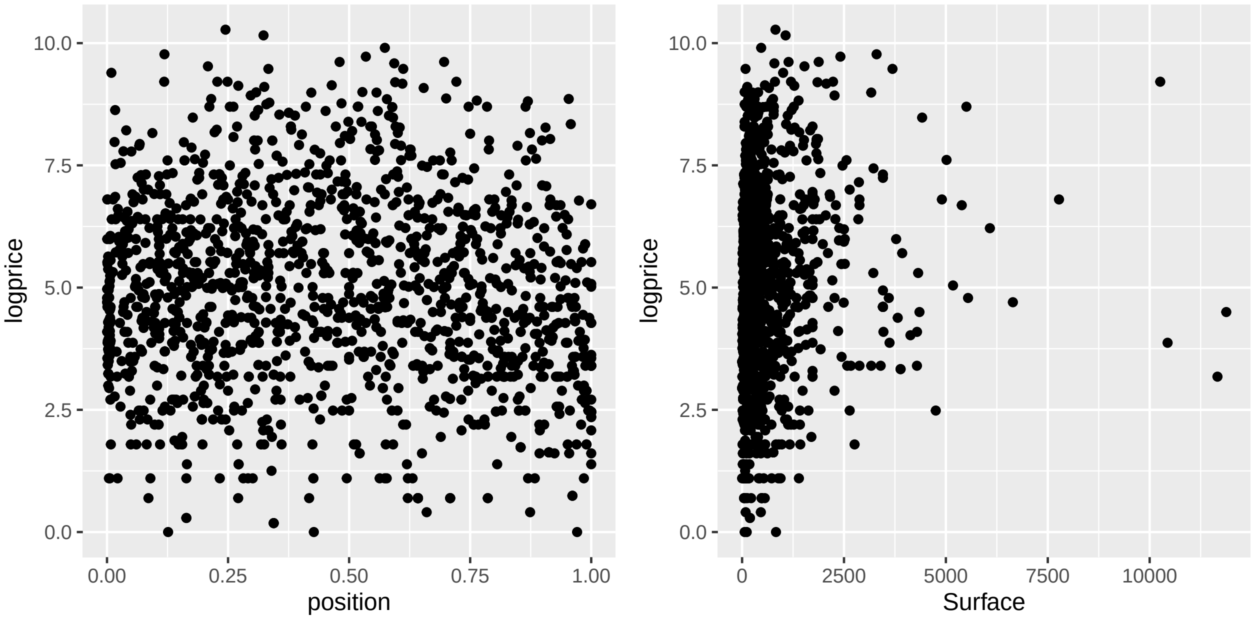


Figure 4: Scatterplots between logprice and position as well as log(Surface)

- The dealers are not dealing each year, every dealer has certain years where they sell paintings, in other years, they don't. This is observed in both the training and the test data.

Variable Exclusion

Further explanation of why certain variables were included/excluded are included in the formal modal analysis below. For brevity, they will not be included here.

Part 3: Discussion of preliminary model Part I

We were fairly happy with our initial model. As we kept removing predictors, every criteria improved (RMSE, coverage, maxDeviation, and MeanAbsDeviation). However, removing predictors has the effect of increasing the bias of our model. We found the bias-variance tradeoff to be worth it, but our model started performing poorly on the training set.

From this, we believe that the training and test data sets are different. We believe this for 2 reasons:

1. The null model outperformed every other model, suggesting that adding predictors made things worse.
2. We used cross-validation to split our training data set into a sub-training and sub-testing data set for several different seeds. The final model fit using the sub-training data performed well on the sub-testing data, suggesting that it should perform well on the test data if the training and testing data were similar. However, we noted much poorer performance on the test data than the sub-test data, again suggesting that the 2 data sets are different.

From these results, we would like to fit a model that performs well on the training data set, then use shrinkage to improve test performance. We need to choose a shrinkage parameter λ that balances fitting the training data well with good test performance.

However, we were later given the correct test set and found our previous model performed much better, so we thought we might not need to constrain the model anymore.

Part 4: Development of the final model

Moving from Initial to Final

Using the old test data

Using the “incorrect” test data set, we followed the same steps as fitting the initial model up to the first step function. Here we noted that the model was performing exceptionally well on the training data compared to the test data. In part I we started to manually remove predictors, but it was difficult to check which order of removal would be optimal. So instead we now opt for Bayesian Model Averaging (BMA) to see if the model can shrink and which variables are unimportant. This also has the added benefit of checking for correlation between the variables.

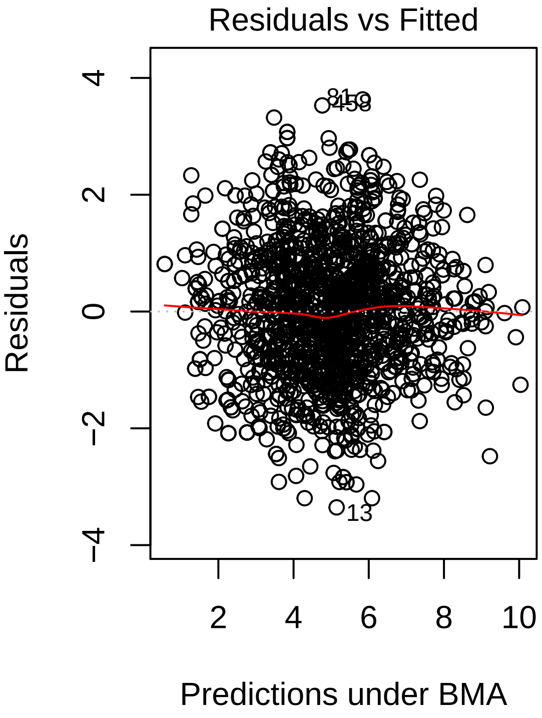
BMA removed a couple variables, but it showed practically every other predictor was associated with the response with high inclusion probability. Additionally, we didn’t find much evidence of correlation between the predictors. The testing performance for this model is still bad, so we will further shrink the model using constraints. Lasso is the technique we used for this.

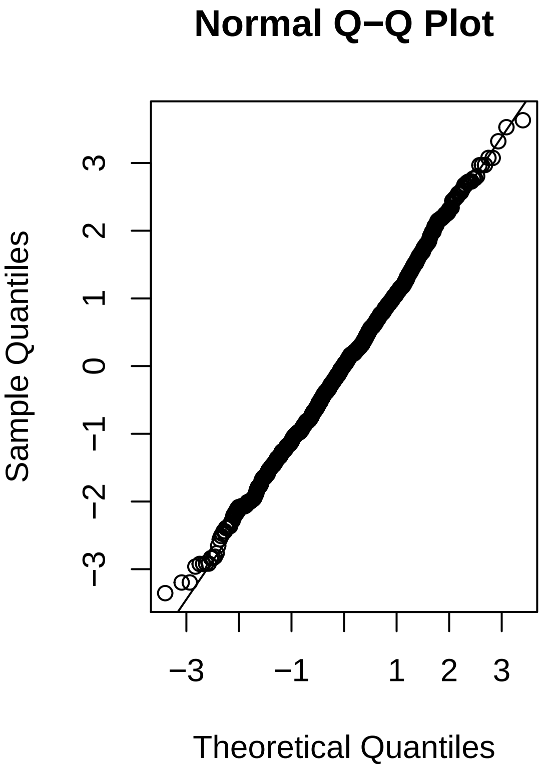
We applied Lasso regression on the best predictive model (BPM), giving an optimal value of $\lambda = 0$ (no shrinkage). As stated previously, this did not result in good performance, so we wanted to manually choose a lambda. After testing, we found $\lambda = 0.3$ to offer similar performances on the training and testing data sets.

Using the updated test data

With the updated test data set, we found that our model started to perform better than the null model and it’s no longer overfitting. So we changed our approach from removing predictors and adding constrains to including more predictors to reduce the bias. We started with initial model again then used a step function but with AIC instead of BIC to gain better prediction performance. Then we used BMA again on the output from the step function and found the model was better than other models we have tried.

Residual Discussion





From these two plots, we found that the residuals of our model were normally distributed around zero and the variance of residuals across fitted values looks constant. We also didn't find any outliers from above plots. It seems that the assumption of normality is satisfied here.

Predictive Intervals

We used the `predict` function to get the standard errors of the predicted values, `se.bma.pred`. Then we manually calculated the 95% prediction interval.

Model Summary

Below we show the coefficient estimates(converted to exponential scale) with their 95% credible intervals and their inclusion probabilities in BMA.

	Fit	2.5%	97.5%	Inclusion Probability
Intercept	130.099	123.159	137.870	1.000
dealerL	5.404	2.861	10.191	1.000
dealerP	2.479	1.333	4.535	1.000
dealerR	3.792	2.664	5.567	1.000
endbuyerB	3.762	2.058	7.130	1.000
endbuyerC	2.511	1.990	3.227	1.000
endbuyerD	2.582	2.117	3.179	1.000
endbuyerE	1.922	1.486	2.509	1.000
endbuyerU	1.574	1.231	1.994	1.000
prevcoll1	2.309	1.751	3.052	1.000
paired1	0.688	0.537	1.040	1.000
finished1	1.786	1.480	2.142	1.000
year1765	5.405	3.596	8.211	1.000
year1766	0.900	0.591	1.381	1.000
year1767	3.760	2.831	4.949	1.000
year1768	1.094	0.777	1.599	1.000
year1769	7.803	4.681	13.145	1.000
year1770	2.978	1.994	4.512	1.000
year1771	3.463	2.414	4.850	1.000
year1772	2.385	1.474	3.816	1.000
year1773	3.231	2.175	4.824	1.000
year1774	5.995	4.202	8.366	1.000
year1775	3.097	1.731	5.417	1.000
year1776	5.421	4.028	7.207	1.000
year1777	11.146	8.139	14.898	1.000
year1778	3.938	2.257	6.858	1.000
year1779	2.074	1.069	3.943	1.000
year1780	2.762	1.497	4.974	1.000
origin_authorD/FL	1.068	0.443	2.543	1.000
origin_authorF	0.729	0.302	1.738	1.000
origin_authorG	0.726	0.287	1.934	1.000
origin_authorI	0.630	0.257	1.510	1.000
origin_authorS	0.993	0.341	3.094	1.000
origin_authorX	0.343	0.147	0.840	1.000
lrgfont1	2.548	2.001	3.228	1.000
engraved1	2.305	1.723	3.016	1.000
diff_origin1	1.539	0.897	2.867	0.999
type_intermedB	2.447	1.219	5.046	0.998
type_intermedD	2.261	1.695	3.077	0.998

	Fit	2.5%	97.5%	Inclusion Probability
type_intermedE	1.717	1.115	2.616	0.998
lands_sc1	0.596	0.474	0.751	0.997
dealerL:diff_origin1	0.930	0.488	1.783	0.966
dealerP:diff_origin1	0.452	0.248	1.000	0.966
dealerR:diff_origin1	0.351	0.188	0.583	0.966
portrait1	0.617	0.448	1.000	0.889
still_life1	0.707	0.487	1.000	0.725
discauth1	1.476	1.000	3.575	0.483
dealerL:paired1	0.705	0.303	1.000	0.370
dealerP:paired1	0.844	0.405	1.012	0.370
dealerR:paired1	0.927	0.625	1.113	0.370
dealerL:discauth1	0.980	0.293	3.369	0.351
dealerP:discauth1	0.368	0.023	1.000	0.351
dealerR:discauth1	0.671	0.235	1.000	0.351
figures1	1.034	1.000	1.350	0.141
othgenre1	1.026	1.000	1.275	0.120
lands_elem1	1.010	1.000	1.112	0.078
peasant1	1.011	0.997	1.092	0.063
type_intermedB:paired1	0.995	1.000	1.000	0.018
type_intermedD:paired1	0.997	1.000	1.000	0.018
type_intermedE:paired1	0.976	1.000	1.000	0.018
type_intermedB:discauth1	1.008	1.000	1.000	0.003
type_intermedD:discauth1	1.004	1.000	1.000	0.003
type_intermedE:discauth1	1.004	1.000	1.000	0.003
type_intermedB:still_life1	0.997	1.000	1.000	0.001
type_intermedD:still_life1	1.004	1.000	1.000	0.001
type_intermedE:still_life1	1.001	1.000	1.000	0.001
dealerL:prevcoll1	1.000	1.000	1.000	0.000
dealerP:prevcoll1	1.002	1.000	1.000	0.000
dealerR:prevcoll1	1.000	1.000	1.000	0.000
type_intermedB:portrait1	0.999	1.000	1.000	0.000
type_intermedD:portrait1	1.000	1.000	1.000	0.000
type_intermedE:portrait1	1.000	1.000	1.000	0.000
endbuyerB:finished1	1.000	1.000	1.000	0.000
endbuyerC:finished1	1.000	1.000	1.000	0.000
endbuyerD:finished1	1.000	1.000	1.000	0.000
endbuyerE:finished1	1.000	1.000	1.000	0.000
endbuyerU:finished1	1.000	1.000	1.000	0.000
endbuyerB:diff_origin1	1.000	1.000	1.000	0.000
endbuyerC:diff_origin1	1.000	1.000	1.000	0.000
endbuyerD:diff_origin1	1.000	1.000	1.000	0.000
endbuyerE:diff_origin1	1.000	1.000	1.000	0.000
endbuyerU:diff_origin1	1.000	1.000	1.000	0.000
endbuyerB:lands_sc1	1.000	1.000	1.000	0.000
endbuyerC:lands_sc1	1.000	1.000	1.000	0.000
endbuyerD:lands_sc1	1.000	1.000	1.000	0.000
endbuyerE:lands_sc1	1.000	1.000	1.000	0.000
endbuyerU:lands_sc1	1.000	1.000	1.000	0.000
type_intermedB:othgenre1	1.000	1.000	1.000	0.000
type_intermedD:othgenre1	1.000	1.000	1.000	0.000
type_intermedE:othgenre1	1.000	1.000	1.000	0.000

From the table above, we see that lots of varaibles have inclusion probability that is close to 1, these predictors are `dealer`, `endbuyer`, `type_intermed`, `origin_author`, `year`, `prevcoll`, `paired`, `finished`, `lrgfont`, `engraved`, `diff_origin`, `lands_sc` and the interaction term `dealer:diff_origin`. Out of these, `paired` and `lands_sc` reduce the price while all the other painting features(the binary predictors) increase it. `portrait` and `still_life` have high probabilities but are not too close to 1 and both seem to reduce the price.
Some predictors, especially all the other interaction terms, have inclusion probability close to 0, and very few predictors have intermediate inclusion probability, suggesting presence of correlation among them.

Since we have a lot of predictors with high inclusion probabilities, we discuss a few of the predictors below which seem to have a large impact on the price of the paintings. We discuss the effects while ignoring the interaction effects, even though most of them do not seem to have any impact(coefficient of 1) anyway.

- For paintings sold in year 1777, we expect the median price in livres increase by 1000% (with a credible interval of (723%, 1420%)), compared to the baseline year 1764. For paintings sold in year 1769, we expect the median price in livres increase by 680% (with a credible interval of (364%, 1237%)), compared to the baseline year 1764. Some of other years also saw an increased price but these two years of sale seemed to be the largest.
- We expect the median price of paintings sold by DealerL in livres increase by 440% (with a credible interval of (192%, 966%)), compared to the baseline dealerJ, when `paired`,`diff_origin`,`discauth`,`prevcoll` are zero. We expect the median price of paintings sold by DealerR in livres increase by 279% (with a credible interval of (163%, 466%)), compared to the baseline dealerJ, when `paired`,`diff_origin`,`discauth`,`prevcoll` are zero. Dealer L supposedly sold the highest priced paintings followed by Dealer R then Dealer P then Dealer J.
- We expect the median price of paintings bought by a buyer in Livres to be 276% (with a credible interval of (99%, 611%)) more than the baseline when the endbuyer is completely unknown(no information, not even the name) when `finished`,`diff_origin`,`land_sc` are zero.
- We expect the median price of paintings with the origin of the author unknown to be 66%(with a credible interval of (19%,86%)) lower than when the author is Austrian. From the results, it seems that Austrian, Dutch/Flemish and Spanish authors' paintings are more expensive.
- We expect the median price of paintings with an additional paragraph included in livres increase by 155% (with a credible interval of (99%, 226%)).
- We expect the median price of paintings in livres to increase by 130% (with a credible interval of (74%, 206%)) if the dealer mentions engravings done after the painting.

We also looked at the posterior densities of the model coefficients. Most of the densities either had a large point mass at 0, or were normally distributed and had a single mode. The few coefficients that were multimodal were `diff_origin`(very small bump), `discauth` and `paired`. Looking into correlations between these and the other predictors may help us improve the model or at least give us some insight into their relationship with price.

Part 5: Assessment of Final Model

Model Evaluation

First, let's get a sense of how much of the variation in price our model is explaining

```
## [1] 0.6647851
```

Our model is explaining approximately 66% of the variation in the log price of paintings. It is important to note that an R^2 cannot be used in a vacuum to tell if the model is good or bad, but it's useful to note. We observe an RMSE of 1538.9, and a coverage of 0.956(using a 95% prediction interval). As stated previously from the residual plot, it looks like all of the assumptions for MLR have been satisfied, so our standard errors + confidence intervals will be appropriate.

Model Testing

For the testing data, we obtained a bias of 220.59. Relative to other groups, this seems like a reasonable bias. We could attempt to include more variables, but we think we already captured the relevant variables associated with log price. Adding more variables here could be beneficial, but likely wouldn't result in too much of a bias improvment. The other statistics here seem reasonable as well:

- Our coverage is 0.956, which is very good
- Our maxDeviation is 13429.63, almost half of some of the other groups. Therefore there are no egregiously different values that our model is predicting.
- Our MeanAbsDeviation is 455.68.
- Finally, our RMSE is 1262.07. For some reason, the RMSE is lower on the test data than the training data. It's possible that this is due to the smaller sample size of the test data. Or it could have "easier" cases to predict.

We believe this model is reasonable due to its large improvement over the null model (no predictors).

Model Result

```
##      dealer                  authorstandard year
## 1001      R Rijn, Rembrandt Harmenszoon van 1769
## 2103      R                 Dujardin, Karel 1776
## 1025      R        Breenbergh, Bartholomeus 1769
## 2478      R Rijn, Rembrandt Harmenszoon van 1777
## 2554      R           Ostade (I), Isack van 1777
## 2617      R              Le Sueur, Eustache 1777
## 1016      R                  Metsu, Gabriel 1769
## 1039      R          Werff, Adriaen van der 1769
## 998       R                       Miel, Jan 1769
## 1042      R     Gell\x8ee (Lorrain), Claude 1769
```

Here are some observations from this:

- The dealer for ALL of the top 10 predicted paintings is R.
- Only one painting was not sold in 1769 or 1777. This observation was only different by one year (1776). These years must have had a good market, or R might have been dealing more than the others.
- One author has 2 paintings in the top 10: Rembrandt Harmenszoon van Rijn. This is the famous Rembrandt, so it's expected. One of the paintings, Arquebusiers / Ronde de nuit, is "The Night Watch", a very famous painting. The other is difficult to find due to the French subject with a incorrectly translated character. It's likely one of the "Holy Family" paintings, a series of 5 famous paintings.
- Most of the endbuyers were collectors that all used dealers as intermediaries.
- All of the top predicted paintings had artists that were either of Dutch/Flemish or French origin. For these, the dealer's catalog also correctly labeled their origin.
- All of these paintings had "lrgfont": the dealer devotes an additional paragraph to each painting

Part 6: Conclusion

Summary of the results

`Dealer`, `Year`, `Endbuyer` and `Origin_author`(when the author is unknown) seem to impact the price of the paintings most(in terms of the coefficient values). We found that the people involved are more important than the painting features, when the people are unknown, the features may become more important though. For example, when the artist in unknown, painting price drops a lot. We also found that certain dealers were more active in certain years. We were also surprised that whether the author was living or not was not important to the price.

If we just want to focus on the features of the paintings and not the people involved, the best features for a painting to have are highly polished finishing, an additional paragraph and engravings done after the painting. The worst features a painting could have are if it is described as a plain landscape, if it is described as a portrait and if its description indicates still life elements. It would be advisable to look for the presence(or absence) of these features to get an idea about the price, however, our model also includes the effects of the people involved in the sale, so we need to be careful if we look at these features on their own. A model to just predict the price of the painting using the features may be better if we only want to look at their effects, but this model would not consider facts such as the reputation of the artist or the dealer etc.

On the old test dataset, we tried various different models including tree models and found that using BMA followed by lasso on the best predictive model gave us a model that could give us reasonable insight into the training data while also not poorly predicting on the test data.

On the new dataset, we used just the BMA model without adding any constraints using lasso. If we had more time, we could have tested tree models like random forests to see if the results are any better. Since, we are just using categorical variables in our model anyway, a tree model might be better suited to this problem in terms of predictions since it can capture many more interactions than a linear model can. The BMA model is still useful as it can give easily interpretable results and we can compare the effects of various predictors.

Things learned

- We should do EDA before we try to build a model and don't forget to consider collinearity bewteen variables.
- Data is not always as clean as the ones we see in homework, we can't always remove observations with missing values. There are some methods to do the missing value imputation, for example, using the MICE package. As far as this project concerned, we found the `Surface` variable had random missing values, but we didn't think we need to impuate the missing values of this variable because it's not very important.
- It's easy to overcomplicated with models, since we have learned lots of techniques. But sometimes keeping it simple may be the best stratagy.