Part 1 Writeup

December 6, 2019

1. Introduction: Summary of problem and objectives (5 points)

We were hired as statistical consultants by an art historian to explore (1) what drove the prices of 18th century paintings in France, and (2) which paintings may be overvalued or undervalued. They have provided us with auction price data from 1764-1780 on the sales (seller/buyer), painter, and other characteristics of paintings.

In part 1, we perform exploratory data analysis to gain high-level insights into the art data, to help us inform further portions, and run a simple model based on these observations. Our objectives of part 1 include:

1. Subsetting data to only consider variables that are both relevant and not redundant of other explanatory variables. For instance, the data contains the variables `material`, `mat`, `materialCat`. These three variables are all conveying similar information, but expressed differently. Which is most appropriate to include in the model?
2. Transforming key text variables to be standardized and in a form that we can pass into regressions and other predictive modelling algorithms. For instance, if we wanted to include the painter (`authorstandard`) in our model, we first may want to remove stop words in the variable such as ("in the taste of"), and may want to count a collaboration between multiple painters (separated by a semicolon in the data) into each individual painter level, rather than count it into a new level of the variable expressing the collaboration separately from the individual painter levels.
3. Impute values of data when they are missing. For instance, the surface area (`Surface`) does not always exist, so we impute the missing data conditional on variables such as `Shape`, `Height_in`, `Width_in`, and `Diam_in` that may give insight into `Surface`.
4. Examining what sources of variation are associated with variations in price for 18th century art sold in France. For instance, given a specific painter, do prices vary significantly? Is the variation for paintings within-painter variation, or across-painter variation?
5. When controlling for all other variables in the dataset, what which variables are most important to predicting the log of the price of paintings. For instance, what is the marginal impact in driving price if it is an Adrien van de Velde painting, holding all other variables constant?
6. Based on the EDA, run an initial model, and examine the in-sample residuals and coverage.

2. Exploratory data analysis (10 points): must include three correctly labeled graphs and an explanation that highlight the most important features that went into your model building.

Part (a) Remove redundancies

There are many redundant variables in the data, so for the purposes of illustration, we show below a sample variable selection choice to remove a redundancy. Note that we used a similar process to remove further redundant variables.

Table 1: Winning Bidder Type vs. End Buyer

		B	C	D	E	U
	395	0	0	0	0	0
B	0	12	0	0	0	0
BB	0	1	0	0	0	0
BC	0	0	10	0	0	0

		B	C	D	E	U
C	0	0	189	0	0	0
D	0	0	0	464	0	0
DB	0	1	0	0	0	0
DC	0	0	89	0	0	0
DD	0	0	0	4	0	0
E	0	0	0	0	127	0
EBC	0	0	1	0	0	0
EC	0	0	37	0	0	0
ED	0	0	0	2	0	0
U	0	0	0	0	0	168

We see that the last digit of `winningbiddertype` always corresponds to the `endbuyer`. We also see, however, that many of the `winningbiddertype` categories are scarcely populated (`EBC`, for instance, only has 1 value in the training data). Therefore, we use `endbuyer`, and exclude `winningbiddertype` after this step.

The following are the variables we excluded (with the variable that conveys similar information in the parentheses) : `origin_author`, `school_pntg`, `diff_origin` (`origin_cat`); `author` (`author_standard`); `winningbiddertype` (`endbuyer`); `type_intermed` (`Interm`); `Surface_Rect`, `Surface_Rnd`, `Height_in`, `Width_in`, `Diam_in` (`Surface`); `materialCat`, `mat` (`material`); `nfigures` (`singlefig`, `figures`); `landsALL` (`lands_*` variables); `lands_ment` (`lands_sc`, `lands_figs`); `lands_elem` (`lands_sc`, `lands_figs`).

Part (b) Cleaning text data

For the purposes of illustration, we only go through the cleaning process for the `authorstandard` variable, though a similar process was applied to `material`, `year`, and `Shape`.

Without any cleaning, it is a good idea to see a set of sample values of the text variable. Below is the 10 most common values for the authors, uncleaned:

Table 2: 10 most Frequent Authors, Uncleaned

authorstandard	total
Teniers II the Younger, David	44
Anonymous	43
Wouwerman, Philips	28
Boucher (F), Franois	26
Rijn, Rembrandt Harmenszoon van	21
La Fosse, Charles de	19
Bourdon, Sbastien	17
French	17
Machy, Pierre-Antoine de	17
Patel I, Pierre	17

To start, we noticed that some `authorstyle`s were included in the `authorstandard` description, where an `n/a` value corresponds to an original by the said author, and all other values correspond to derivations of the author's original work:

Table 3: Types of Derivations of Painters' Work

Author Style	Frequency
after	26

Author Style	Frequency
attributed to	7
copy after	10
esteem of	1
in the genre of	3
in the manner of	7
in the style of	7
in the taste	1
in the taste of	17
n/a	1417
school of	2
sketch of	1
taste of	1

For the latter group, we marked all such rows as a copy (by creating a variable `is_copy`), and removed these keywords from the author. Additionally, there are some instances where multiple authors contributed to a painting. As a result, we created a binary variable for each other who contributed to at least 10 paintings, where a 1 indicates contribution. Below, observe the sample counts for the most frequent set of authors:

Table 4: 10 most Frequent Authors, Cleaned

authorstandard	total
teniers_ii_the_younger_david	51
anonymous	43
wouwerman_philips	35
boucher_f_fran_u_fffd_ois	30
breughel_i_the_elder_j_jan	26
rijn_rembrandt_harmenszoon_van	24
la_fosse_charles_de	19
patel_i_pierre	19
machy_pierre_antoine_de	18
poelenberch_cornelis_van	18

While a lot of the most frequent are the same, we see that some, such as Jan Breughel the Elder, now appear in the data (he contributed with many others in the dataset), and others have an increased count, such as David Teniers the Younger.

Part (c) Imputing missing data

It is a good idea to examine which variables contain missing data, and how frequently:

Table 5: Missing Variables

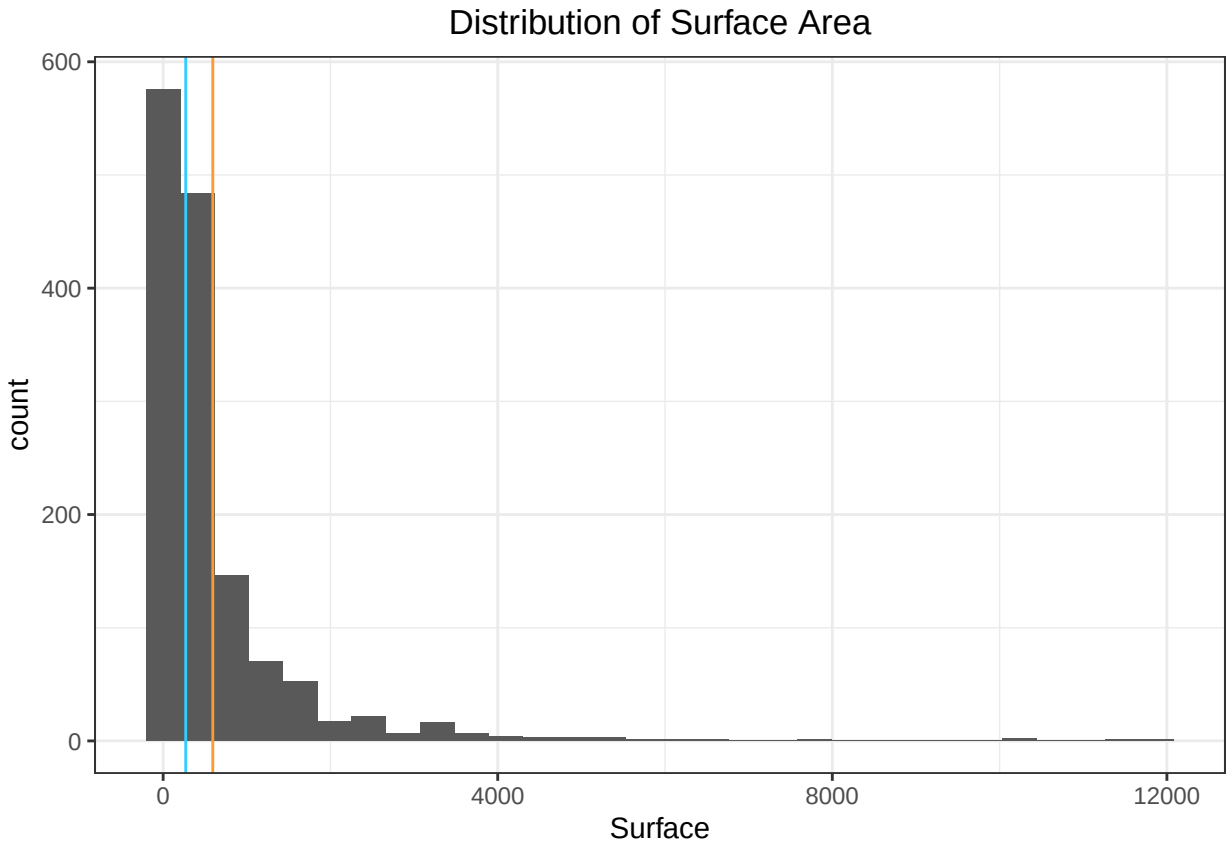
Variables	Missing.Rate
Diam_in	0.979
authorstyle	0.945
Surface_Rnd	0.916
type_intermed	0.903
winningbidder	0.263
winningbiddertype	0.263
endbuyer	0.263

Variables	Missing.Rate
Interm	0.263
materialCat	0.127
mat	0.095
Width_in	0.077
Surface_Rect	0.077
Height_in	0.075
material	0.073
Surface	0.054
Shape	0.013

Since we already removed many of these variables from consideriation in the data redundancy step, we only need to impute values of `Interm`, `endbuyer`, `Surface`.

Since `Interm` and `endbuyer` are categorical, we created a separate category for the missing data.

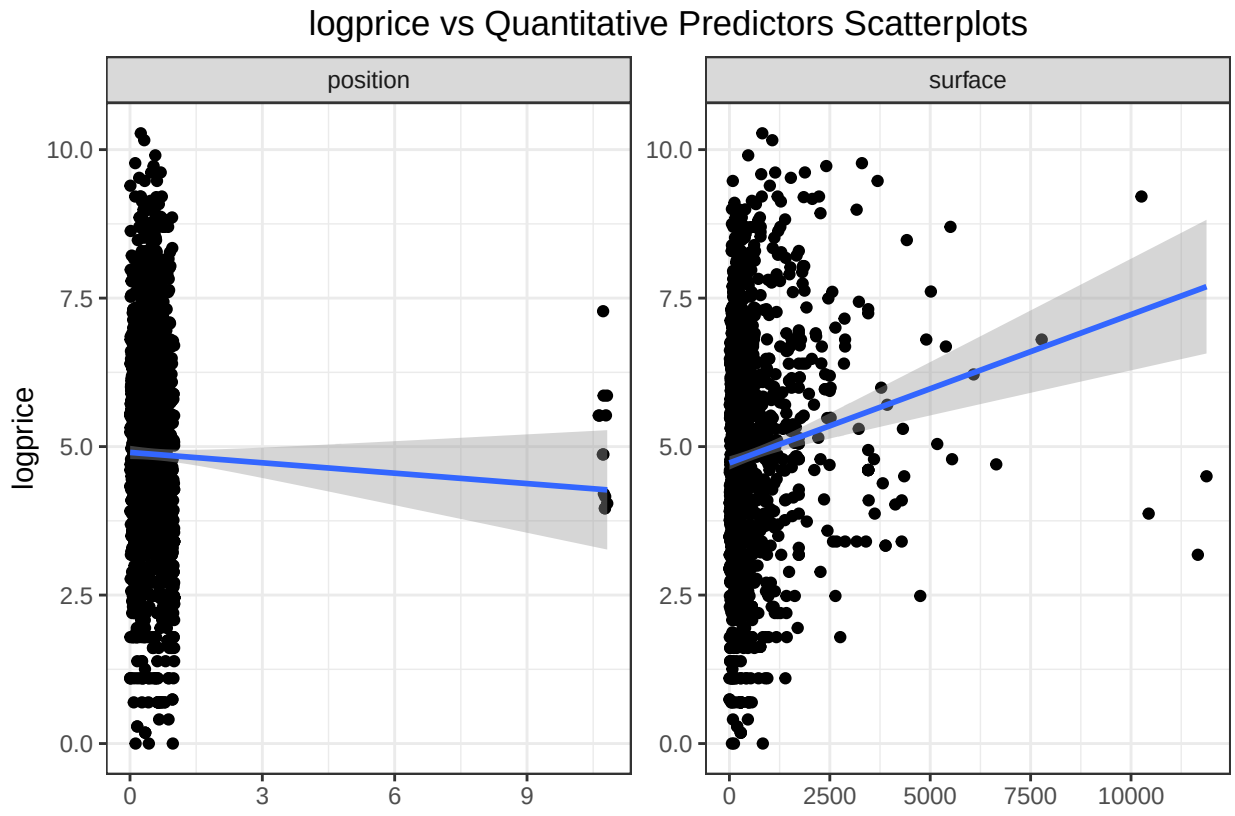
Since `Surface` is continuous, we used median imputation. We chose median imputation because the median is less sensitive to outliers than the mean, and we can see that the data is skewed (blue line represents the median, orange line represents the mean):



Additionally, we noticed that some covariates were not missing when `Surface` is missing, such as `Shape`, `Height_in`, `Width_in`, and `Diam_in`, so we conditioned our imputed values on the non-missing of these variables. For instance, observation 745 is a square rectangle painting with a height of 17 inches, so our imputed value for the surface area is 340 sq. inches, instead of the imputed 283.5 sq. inches for a square rectangle painting with no height or width filled in.

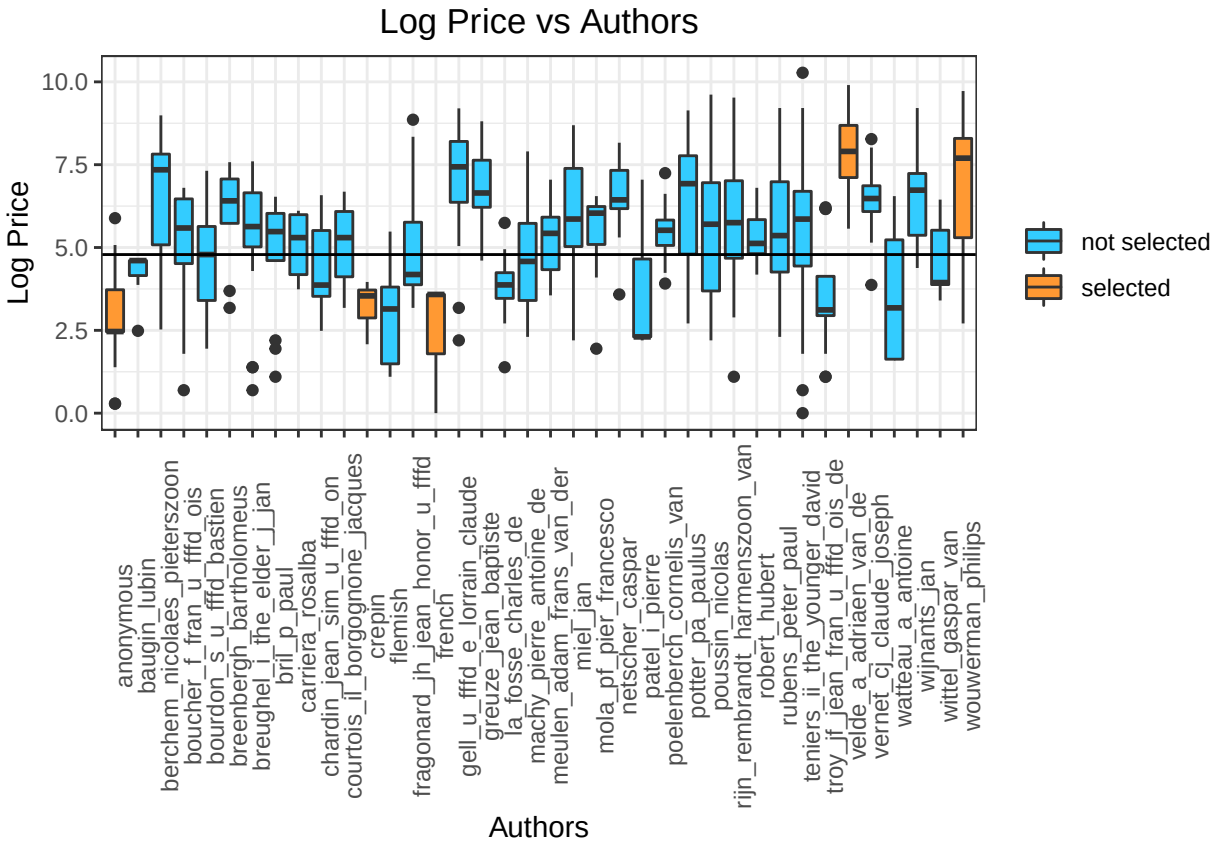
Part (d) Variation of variables

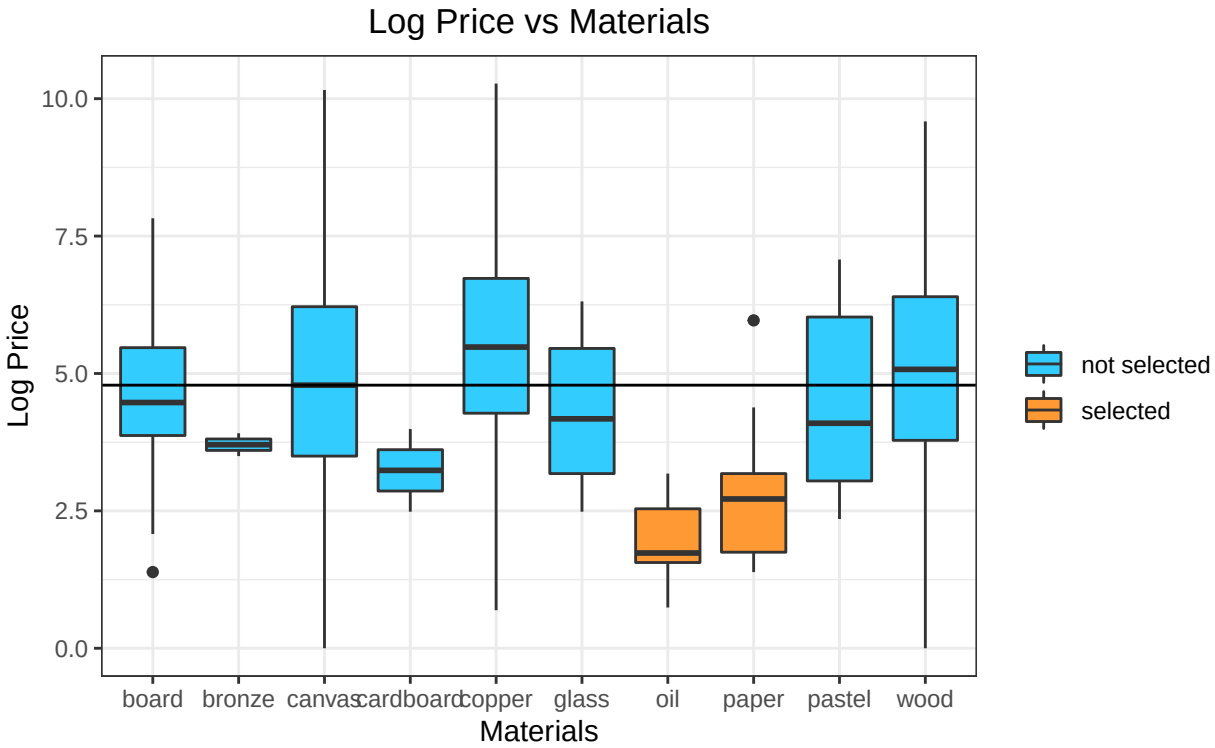
We start by examining the continuous variables:

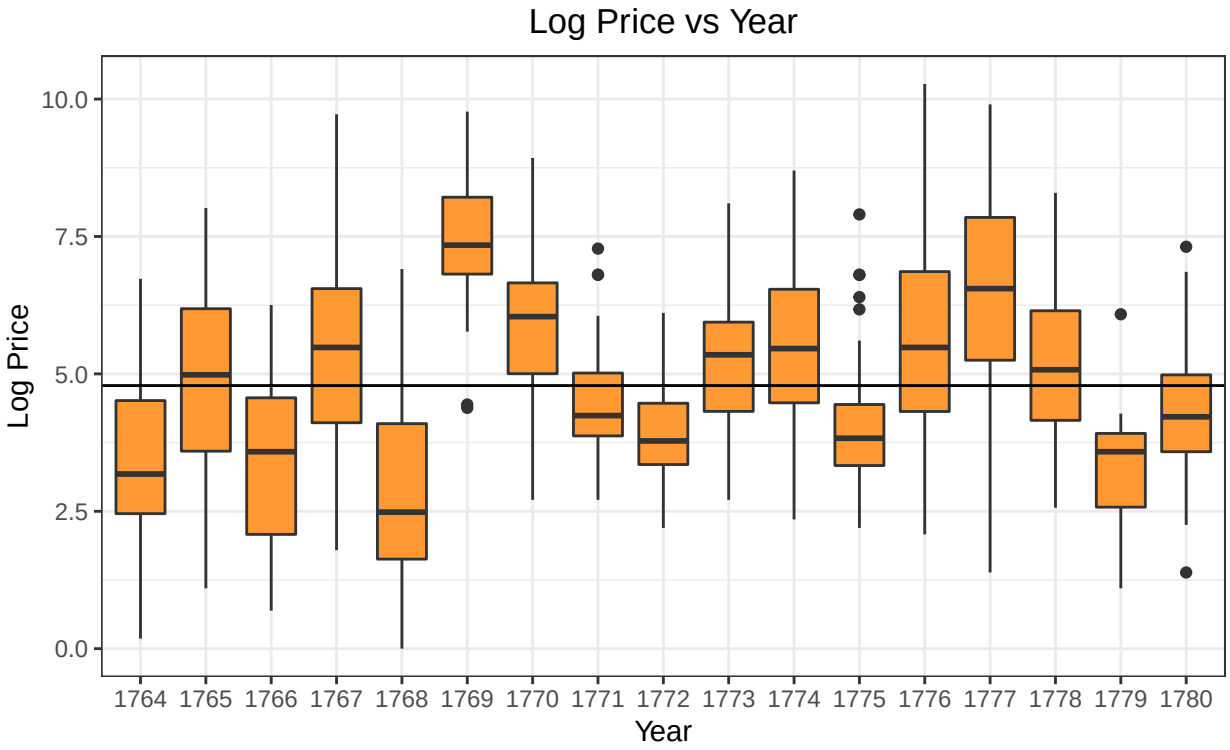


We see above that position doesn't seem to be strongly related to price, whereas there may be some relationship between surface and price.

Next, we consider categorical variables. Specifically, we want to examine the sources of variation of key discrete variables (within-variable variation or between-variable variation). We look at painters who appear at least 10 times in the dataset, the material of the painting, and the year the painting was sold:


Log Price vs Authors
Log Price
10.0
7.5
5.0
2.5
0.0
anonymous
baugin_lubin
berchem_nicolaes_pieterszoon
boucher_f_fran_u_fffd_ois
bourdon_s_u_fffd_bastien
breenbergh_bartholomeus
breughel_i_the_elder_j_jan
bril_p_paul
carriera_rosalba
chardin_jean_sim_u_fffd_on
courtois_il_borgognone_jacques
crepin
flemish
fragonard_jh_jean_honor_u_fffd
french
gell_u_fffd_e_lorrain_claude
greuze_jean_baptiste
la_fosse_charles_de
machy_pierre_antoine_de
meulen_adam_frans_van_der
miel_jan
mola_pf_pier_francesco
netscher_caspar
patel_i_pierre
poelenberch_cornelis_van
potter_pa_paulus
poussin_nicolas
rijn_rembrandt_harmenszoon_van
robert_hubert
rubens_peter_paul
teniers_ii_the_younger_david
troy_jf_jean_fran_u_fffd_ois_de
velde_a_adriaen_van_de
vernet_cj_claude_joseph
watteau_a_antoine
wijnants_jan
wittel_gaspar_van
wouwerman_philips
Authors
not selected
selected


Log Price vs Materials
Log Price
10.0
7.5
5.0
2.5
0.0
board
bronze
canvas
cardboard
copper
glass
oil
paper
pastel
wood
Materials
not selected
selected



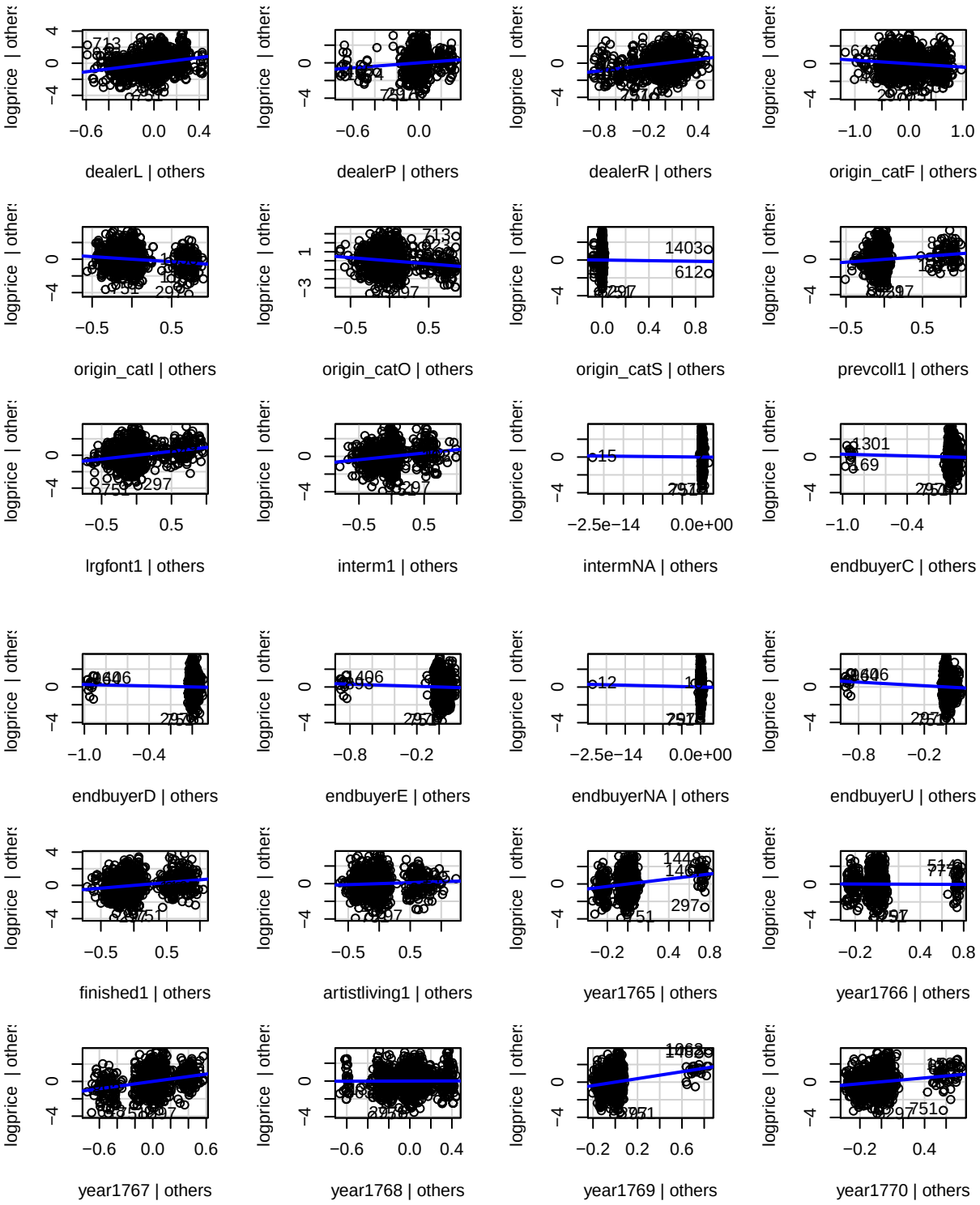
We can see that there exists both within-class and between-class variation for all three variables, indicating that we may want to include some of the individual levels of the variables (we highlighted the levels that we eventually use in the model in part (3) in orange), but will likely also want to control for other sources of variation.

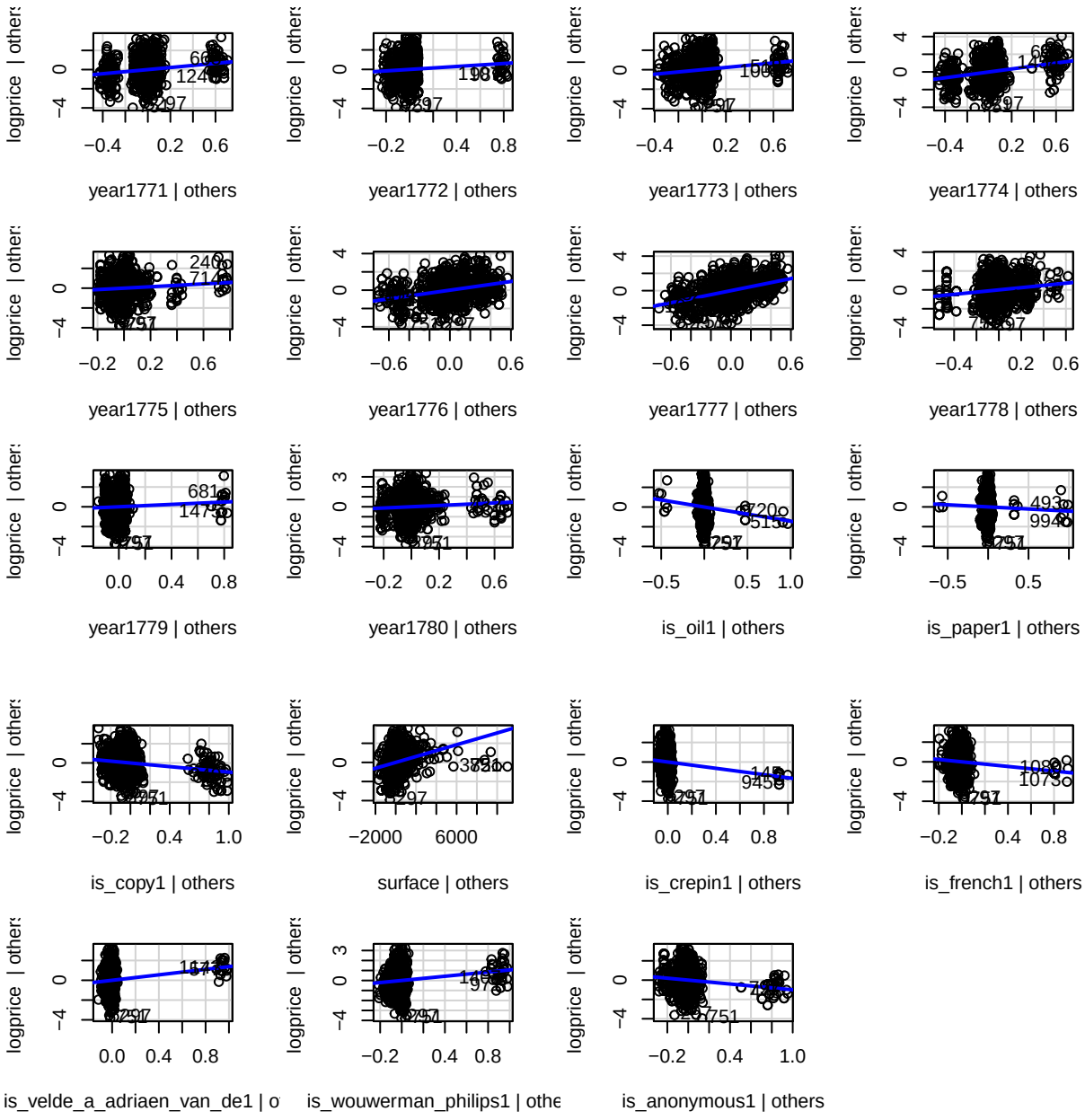
A final part we wanted to consider in sources of variation is in potential interactions. Specifically, we thought that dealers may specialize in different kinds of art, and therefore have different variation in prices for different kinds of paintings. Below, we consider the dealer's prices conditional on the origin of the work (`origin_cat`):

We can see that dealer "R" is the exclusive dealer in Spanish paintings in this dataset, and they sold for much more than the other pieces he sold. Additionally, it looks like dealer "J" sold a small amount of highly priced works of non-Spanish, non-Dutch, non-French, and non-Italian descent, while the other dealers did not, indicating that perhaps dealer "J" had exclusive access to a specific painter/set of painters outside of western Europe.

Part (e) AV-plots

As we saw in part (e), looking at individual boxplots shows some but not all of the variation of key categorical variables. We wanted to further this analysis by robustly considering the marginal impact a variable has holding all other variables constant. Below, we plot add-variable plots of all variables we consider for our model:


logprice | others
dealerL | others
dealerP | others
dealerR | others
origin_catF | others
origin_catI | others
origin_catO | others
origin_catS | others
prevcoll1 | others
lrgfont1 | others
interm1 | others
intermNA | others
endbuyerC | others
endbuyerD | others
endbuyerE | others
endbuyerNA | others
endbuyerU | others
finished1 | others
artistliving1 | others
year1765 | others
year1766 | others
year1767 | others
year1768 | others
year1769 | others
year1770 | others

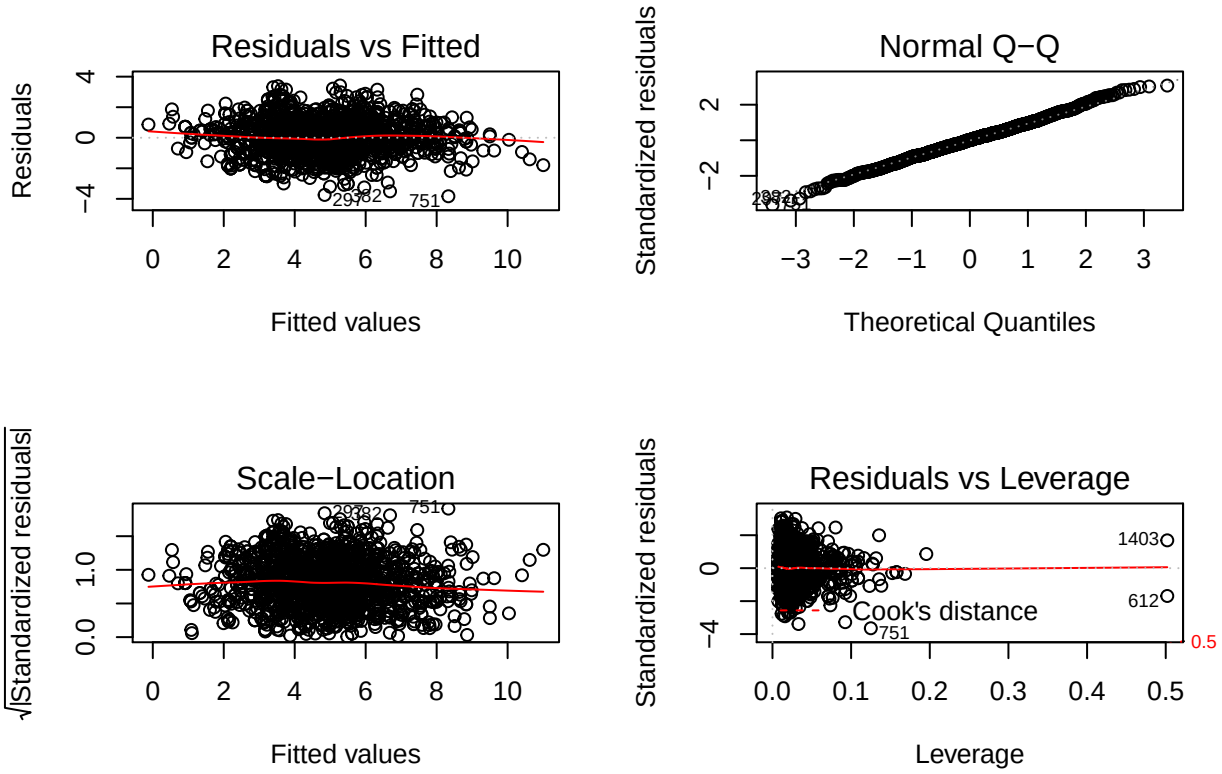


3. Development and assessment of an initial model (10 points)

Part (a) Initial model: must include a summary table and an explanation/discussion for variable selection and overall amount of variation explained.

Based on our EDA, we wanted to include variables that had significant slopes in the AV-plots, since once controlling for other variables, these variables still are associated with the unexplained portion of `logprice`. This led us to choose `dealer`, `origin_cat`, `prevcoll`, `lrgfont`, `interm`, `endbuyer`, `finished`, `artistliving`, `surface`, `year` (as a factor), `is_copy` (created from `authorstyle`), `is_oil` (created from `material`), `is_paper` (created from `material`), `is_crepin` (created from `authorstandard`), `is_french` (created from `authorstandard`), `is_velde_a_adriaen_van_de` (created from `authorstandard`), `is_wouwerman_philips` (created from `authorstandard`), and `is_anonymous` (created from `authorstandard`) in our initial model:

```
## lm(formula = logprice ~ dealer + origin_cat + prevcoll + lrgfont +
##     interm + endbuyer + finished + artistliving + surface + factor(year) +
##     is_oil + is_copy + is_paper + is_crepin + is_french + is_velde_a_adriaen_van_de +
##     is_wouwerman_philips + is_anonymous, data = paintings_train2)
```



We can see that the residuals are pretty consistently centered around zero across different values of the predicted `logprice`, appear generally normal, and are not especially heteroscedastic.

Table 6: Confidence interval of coefficients from initial model

Coefficient	Estimate	2.5%	97.5%
(Intercept)	3.1168	2.4352	3.7985
factor(year)1777	2.3286	2.0507	2.6066
factor(year)1769	1.9719	1.4574	2.4864
dealerL	1.7685	1.4148	2.1223
factor(year)1774	1.6706	1.3354	2.0057
factor(year)1776	1.5775	1.3092	1.8459
factor(year)1765	1.4334	1.0328	1.8339
factor(year)1767	1.4012	1.1272	1.6752

Coefficient	Estimate	2.5%	97.5%
is_velde_a_adriaen_van_de1	1.3783	0.7934	1.9632
factor(year)1778	1.1787	0.7931	1.5642
factor(year)1773	1.1603	0.7696	1.5509
dealerR	1.1278	0.8744	1.3812
factor(year)1770	1.0864	0.6930	1.4797
factor(year)1771	1.0828	0.7465	1.4190
is_wouwerman_philips1	1.0565	0.6640	1.4491
lrgfont1	0.9408	0.7080	1.1736
dealerP	0.9384	0.4262	1.4506
factor(year)1772	0.7581	0.2871	1.2291
interm1	0.7534	0.4999	1.0069
factor(year)1775	0.7178	0.1973	1.2383
finished1	0.6715	0.4986	0.8444
prevcoll1	0.6432	0.3691	0.9172
factor(year)1780	0.6339	0.1140	1.1537
factor(year)1779	0.5983	-0.0435	1.2401
artistliving1	0.2229	0.0293	0.4166
factor(year)1768	0.0314	-0.2782	0.3410
surface	0.0003	0.0002	0.0004
factor(year)1766	-0.0440	-0.4651	0.3771
origin_catS	-0.1797	-1.8119	1.4525
endbuyerD	-0.2727	-0.8804	0.3350
endbuyerC	-0.3139	-0.9253	0.2975
origin_catF	-0.3883	-0.5496	-0.2269
endbuyerE	-0.3946	-1.0306	0.2413
is_paper1	-0.4266	-1.1653	0.3120
origin_catI	-0.6151	-0.8187	-0.4116
endbuyerU	-0.6765	-1.2995	-0.0536
origin_catO	-0.6932	-0.9343	-0.4520
is_copy1	-0.9506	-1.2197	-0.6814
is_anonymous1	-1.0017	-1.3798	-0.6237
intermNA	-1.1312	-1.7535	-0.5089
is_french1	-1.1744	-1.7628	-0.5860
is_oil1	-1.4354	-2.2700	-0.6007
is_crepin1	-1.6129	-2.3340	-0.8917
endbuyerNA	NA	NA	NA

Table 7: R2 of the Initial Model

r.squared
0.6668517

We can see that the initial model explains roughly 2/3 of the variation in `logprice` for the training data, and the only variables whose CIs for the coefficient estimates cross zero are individual levels of some factor variables, indicating that all of the variables chosen in the initial model help explain what drove prices of paintings. However, in a more refined search, it may be beneficial to remove certain levels of some variables that are not appearing as valuable (perhaps we can include binary versions of the `years` variable for each year from 1764-1780 and select the significant ones instead of forcing each year to have a coefficient estimate).

Part (b) Model selection: must include a discussion

We first tried to modify our model by adding interactions, and thereafter used stepwise selection to further refine our model. We enumerated all interactions initially, and below provide a formula for a model with interactions that were significant enough by deviance:

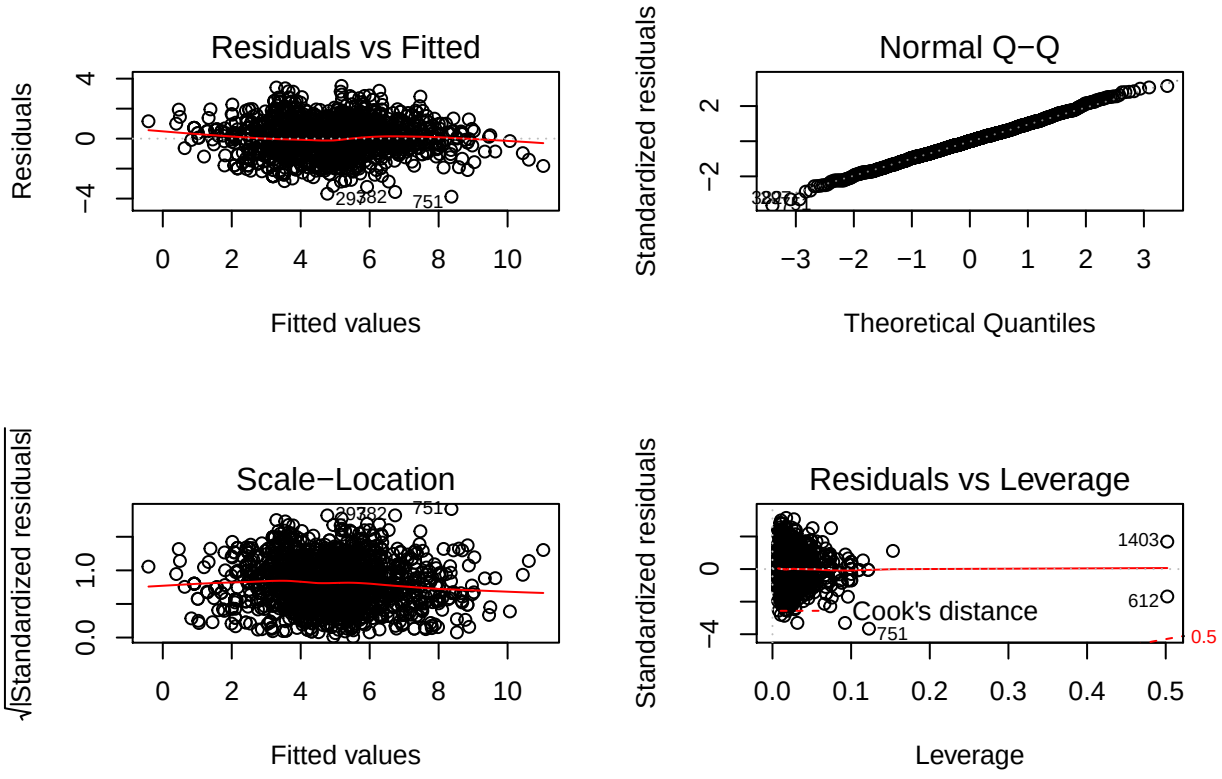
```
## logprice ~ dealer + origin_cat + prevcoll + lrgfont + interm +
##     endbuyer + finished + artistliving + surface + year + is_oil +
##     is_copy + is_paper + is_crepin + is_french + is_velde_a_adriaen_van_de +
##     is_wouwerman_philips + is_anonymous + dealer:origin_cat +
##     dealer:artistliving + dealer:is_paper
```

As we may have expected based on our EDA, dealer interactions could be valuable in modelling art prices.

Since we introduce new sources of variation, we decide to use stepwise selection with a BIC penalty to further refine the model before assessing goodness of fit and coverage. We considered both AIC and BIC models, but chose the BIC model, as we are not only trying to predict overvalued/undervalued paintings in this problem, but trying to explain what drives prices. Since BIC tries to find the true model (rather than the model that best explains the unknown), this more naturally fits our problem of explaining what drove prices.

Part (c) Residual: must include residual plot(s) and a discussion.

Below is a residual plot for our BIC-selected model:



We can see that similar to our initial model, the residuals are generally centered around zero across all values of predicted `logprice` (though even stabler for the BIC-selected model), are generally normally, and are generally homoscedastic.

Part (d) Variables: must include table of coefficients and CI

Table 8: Best Model Summary Table With C.I. of the Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)	lwr	upr
(Intercept)	2.681	0.164	16.366	0.000	2.360	3.003
dealerL	1.805	0.181	9.997	0.000	1.451	2.159
dealerP	0.921	0.261	3.529	0.000	0.409	1.433
dealerR	1.191	0.128	9.285	0.000	0.939	1.442
origin_catF	-0.336	0.077	-4.365	0.000	-0.487	-0.185
origin_catI	-0.635	0.104	-6.101	0.000	-0.839	-0.431
origin_catO	-0.676	0.123	-5.505	0.000	-0.916	-0.435
origin_catS	-0.202	0.837	-0.242	0.809	-1.844	1.439
prevcoll1	0.644	0.140	4.594	0.000	0.369	0.919
lrgfont1	0.946	0.119	7.944	0.000	0.713	1.180
interm1	0.767	0.113	6.783	0.000	0.545	0.989
intermNA	-0.760	0.094	-8.086	0.000	-0.945	-0.576
finished1	0.684	0.088	7.752	0.000	0.511	0.858
surface	0.000	0.000	9.626	0.000	0.000	0.000
year1765	1.482	0.204	7.263	0.000	1.082	1.882
year1766	-0.055	0.214	-0.257	0.798	-0.474	0.365
year1767	1.429	0.139	10.289	0.000	1.156	1.701
year1768	0.060	0.157	0.380	0.704	-0.248	0.367
year1769	2.020	0.262	7.696	0.000	1.505	2.535
year1770	1.140	0.199	5.723	0.000	0.749	1.530
year1771	1.088	0.171	6.375	0.000	0.753	1.423
year1772	0.777	0.240	3.240	0.001	0.306	1.247
year1773	1.209	0.198	6.121	0.000	0.822	1.597
year1774	1.732	0.170	10.189	0.000	1.398	2.065
year1775	0.793	0.265	2.989	0.003	0.273	1.314
year1776	1.631	0.133	12.301	0.000	1.371	1.891
year1777	2.380	0.139	17.074	0.000	2.107	2.654
year1778	1.232	0.197	6.256	0.000	0.846	1.619
year1779	0.568	0.328	1.731	0.084	-0.075	1.211
year1780	0.688	0.264	2.604	0.009	0.170	1.206
is_oil1	-1.577	0.359	-4.396	0.000	-2.280	-0.873
is_copy1	-0.971	0.138	-7.043	0.000	-1.241	-0.700
is_crepin1	-1.683	0.369	-4.560	0.000	-2.407	-0.959
is_french1	-1.244	0.289	-4.299	0.000	-1.812	-0.677
is_velde_a_adriaen_van_de1	1.394	0.299	4.663	0.000	0.808	1.981
is_wouwerman_philips1	1.050	0.201	5.223	0.000	0.656	1.445
is_anonymous1	-1.035	0.193	-5.373	0.000	-1.413	-0.657

Table 9: R2 of the Best Model

r.squared
0.6614811

We see that the direction and magnitude of the variables included has not changed much compared to the initial models, though we have dropped `endbuyer`, `artistliving`, and `is_paper`. As a result, the CIs similarly only cross zero for a few levels of individual factors (specifically for `year`), and the in-sample

R-squared ever-so-slightly decreased due to including fewer variables.

4. Summary and Conclusions (10 points)

What is the (median) price for the "baseline" category if there are categorical or dummy variables in the model (add CI's)? (be sure to include units!) Highlight important findings and potential limitations of your model. Does it appear that interactions are important? What are the most important variables and/or interactions? Provide interprations of how the most important variables influence the (median) price giving a range (CI). Correct interpretation of coefficients for the log model desirable for full points.

Provide recommendations for the art historian about features or combination of features to look for to find the most valuable paintings.

Answer:

From the summary table of our best model, it seems like BIC takes out all the interactions we selected. The following is a list of included categorical variables: dealer, origin_cat, prevcoll, lrgfont, interm, finished, year, is_oil, is_copy, is_crepin, is_french, is_velde_a_adriaen_van_de, is_wouwerman_philips, is_anonymous. The baseline is the painting sold by dealer J in 1764, where the previous owner is not mentioned, the dealer did not devote an additional paragraph, an intermediary is not involved, origin of painting is Dutch/Flemish, the painting is not noted for its highly polished finishing, the material is not oil, is not a copy, and is not created by Crepin, French, Velde A Adriaen Van De, Wouwerman Philips, or anonymous author. The baseline painting price can be estimated by $exp(2.681 + 0.0003surface)$.

Let us check the most important variables and their corresponding confidence intervals:

Table 10: Most important variables

var	coef	lwr	upr
dealerL	1.8046	1.4505	2.1587
dealerR	1.1907	0.9392	1.4423
surface	0.0003	0.0002	0.0004
year1767	1.4286	1.1562	1.7010
year1774	1.7318	1.3984	2.0652
year1776	1.6312	1.3711	1.8914
year1777	2.3804	2.1070	2.6539

From the table above, we noticed that dealer, surface, and year are the three most important variables that influence the log price of paintings. The interpretations are as following:

For variable dealer, the baseline is dealer J. Keeping all other variables constant, we would expect the price of a painting sold by dealer L is exp(1.8046) =6.078 times of the price of dealer J. The confidence interval is 1.4505 to 2.1587, which means we are 95% confident that dealer L's price is exp(1.4505) = 4.265 to exp(2.1587) = 8.66 times of the price of dealer J. Similarly, the price of a painting sold by dealer R is exp(1.1907) = 3.289 times of the price of dealer J, and we are 95% confident that dealer L's price is exp(0.9392) = 2.558 to exp(1.4423) = 4.23 times of the price of dealer J.

For surface, keeping all other variables constant, if the surface increased by 1 squared inches, we would expect the price of the painting to increase exp(0.0003)-1 = 0.03%, and we are 95% confident that the increase will be between exp(0.0002)-1 = 0.02% and exp(0.0004)-1 = 0.04%.

For variable year, the baseline is year 1764. Therefore, if we keep all other variables constant, we would expect the painting price in 1767, 1774, 1776, 1777 is exp(1.4286) =4.173, exp(1.7318) =5.651, exp(1.6312) =5.11, exp(2.3804) =10.809 times of the price in 1764, correspondingly. Also, we are 95% confident that the painting price in 1767, 1774, 1776, 1777 are 3.178 – 5.479,4.049 – 7.887, 3.94 – 6.629, 8.224 – 14.209 times of the painting price in 1764, correspondingly.

Then, let us consider the median, the 2.5th percentile, and the 97.5th percentile for our predictions vs. the actual data:

Table 11: Predicted vs. Actual Quantiles

	50th percentile	2.5th percentile	97.5th percentile
Predicted Prices	140.63 livres	7.36 livres	3396.52 livres
Actual Prices	120 livres	3 livres	6000.52 livres

We can see that the actual data has much larger values at the extreme than our models predict. It could be because sometimes, if a rich person really likes a painting, they could be a victim of the winner's curse (see: https://www.investopedia.com/terms/w/winnerscurse.asp).

Additionally, the predicted median is higher than the actual median. This could be because the model was slightly influenced by the points that could have been victim to the winner's curse.

Now, let's consider coverage:

Table 12: Coverage of the training data

x
95.0%

We see that our model is very well calibrated on the training data, where 95% of our data fall within the 95% prediction interval. While we were off in middle and at the highly priced items in the median prediction, we see that we do a fairly good job accounting for uncertainty in the training data.

While the coverage in the training data is encouraging, it is worth considering shortcomings of our model. Firstly, we only considered a subset of the painters in the model, due to having them each be binary variables and the selection methods only including a few (this is a function of some paintings having multiple painters, so there are shortcomings with encoding `authorstandard` as a multi-class categorical variable and including it once in the model). However, we may want to think of artwork prices as being part of a hierarchical model: different painters have an inherent skill level, drawn from a distribution, and each individual work of theirs has a certain quality, drawn from a distribution based on the painter's skill. For instance, an Adrien van de Velde work is on average worth 4.0328207 that of any other random painting, holding all other variables constant, with a 95% CI of (2.2430134, 7.2508003). On the other hand, if the painter is anonymous, his work is on average worth 0.3552492 that of any other random painting, holding all other variables constant, with a 95% CI of (0.2434716, 0.5183437). Therefore, it may be worthwhile to include more levels for painters in the future, and design a model that treats the data more like a hierarchical model. Buyers may be more interested in a painting if they heard of or respected the artist, regardless of the theme of the piece, the size of it, or what materials were used. It is widely known, for instance, that the supposed last Leonardo Da Vinci painting sold for an extremely large amount of money (see: https://news.artnet.com/market/last-known-leonardo-da-vinci-painting-just-sold-1149032).

It may be useful to also consider on the dealer (and maybe even the buyer!), which will be an even more complicated interaction hierarchy (and would require cleaning the buyer names as well). Once we are already trying to robustly control for one agent in the art auction process (the artist), why not account for the other agents?

Finally, below we present a few undervalued pieces of artwork based on our model, followed by overvalued pieces of artwork:

Table 13: Undervalued Paintings

Sale (Dealer+Year of Sale)	Artist	Actual Price	Predicted Price
R1777	Rubens, Peter Paul	10000.0	62005.677
R1777	Wijnants, Jan; Velde (A), Adriaen van de	10000.0	40770.291
R1777	Lairesse, Grard de	13001.0	34269.453
R1776	Wouwerman, Philips	6520.0	15413.663
R1769	Breenbergh, Bartholomeus	1750.5	8404.286

It looks like that dealer "R" in 1776 and 1777 may have been undervaluing some of his paintings, specifically from renowned artists such as Philips Wouwerman.

Table 14: Overvalued Paintings

Sale (Dealer+Year of Sale)	Artist	Actual Price	Predicted Price
R1776	Teniers II the Younger, David	29000	1756.6610
R1776	Metsu, Gabriel	25800	7163.8700
R1777	Poussin, Nicolas	15000	1715.6031
R1769	Murillo, Bartolom Esteban	17535	4580.2897
R1777	Dou, Gerrit	13000	533.1464

For the overvalued paintings, this could partially be a function of our current model. We notice that none of the artists who had the most overvalued paintings were directly represented in our model, indicating that perhaps their work was very appreciated, but we did not capture it in our linear regression. It is interesting, however, that these overvalued paintings all came from the same dealer as the undervalued paintings (during a similar time period too!), which may mean he was selling to a different clientele than the rest of the dealers in a way that our best linear regression model cannot recognize.