Part I: Simple Model

Linlin Li, Bingruo Wu, Cole Juracek, Vidvat Ramachandran

06 December, 2019

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.

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.

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

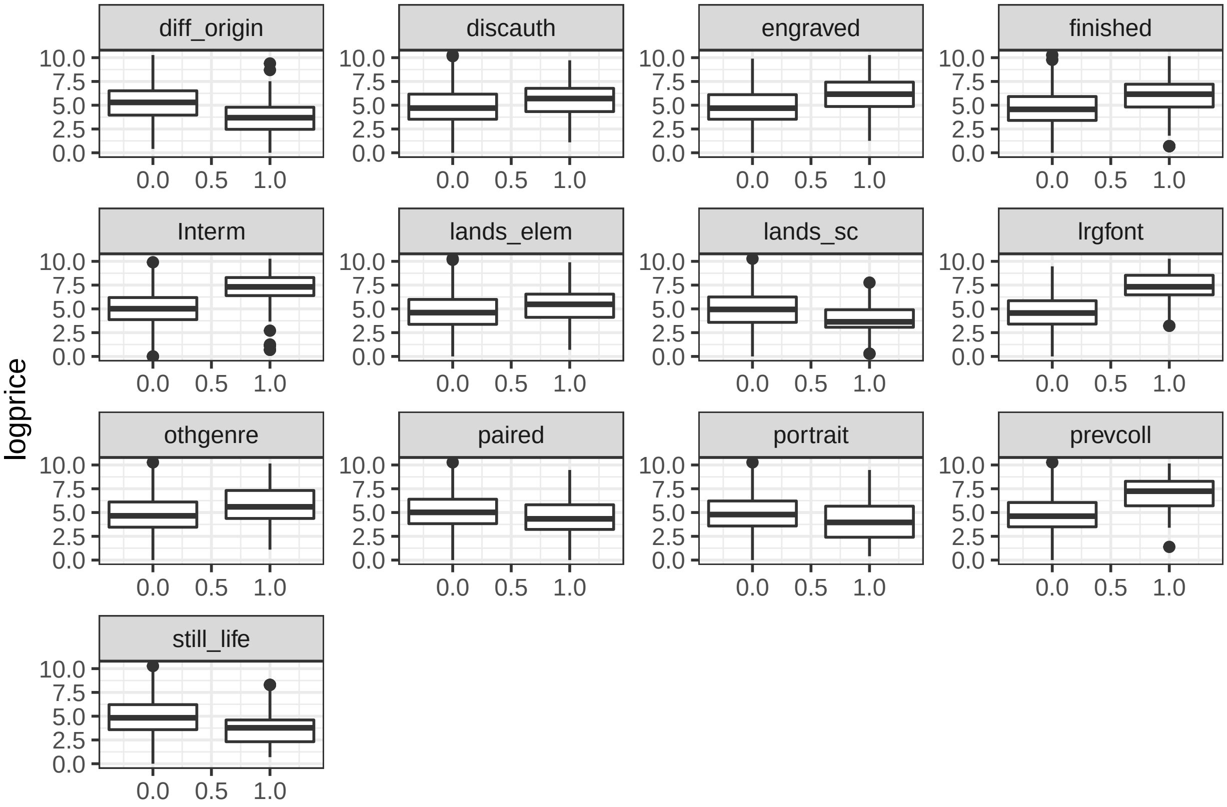

Figure 1: Boxplots between logprice and some binary variables

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.

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.

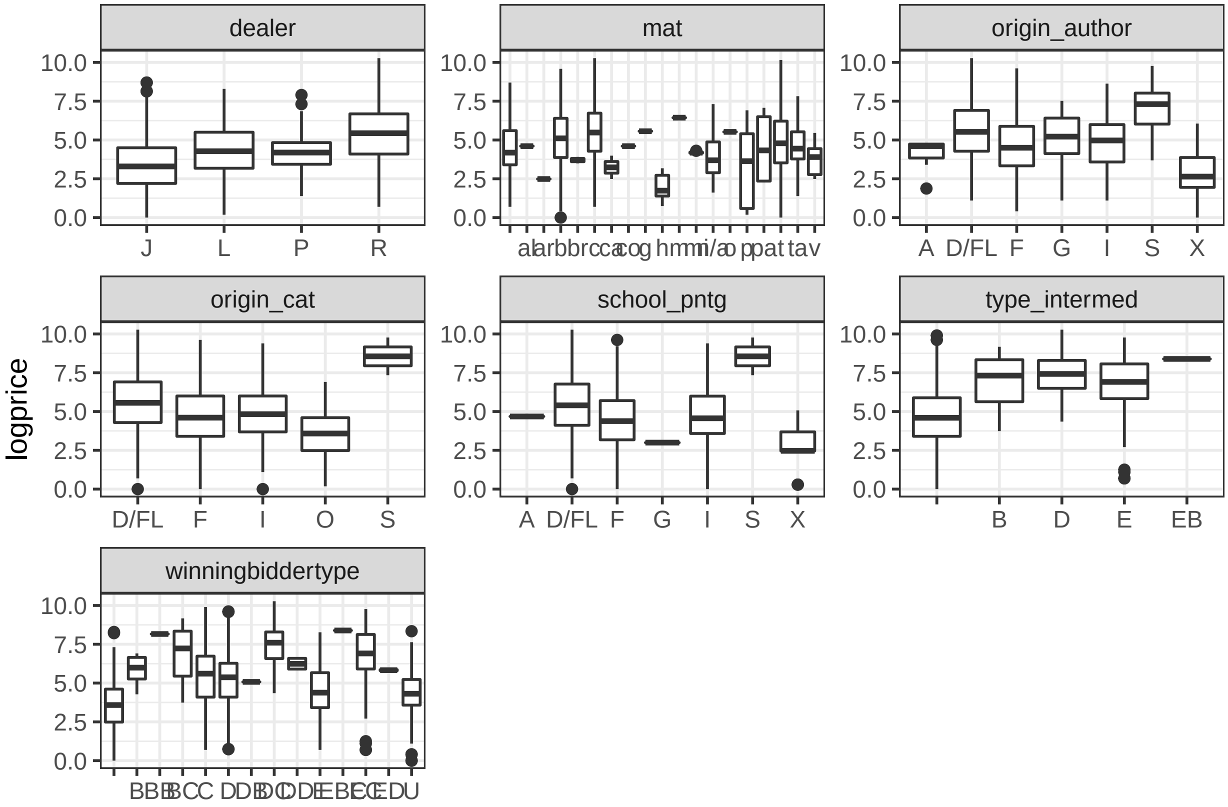

Figure 2: Boxplots between logprice and some categorical variables

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`.

Continuous Variables

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

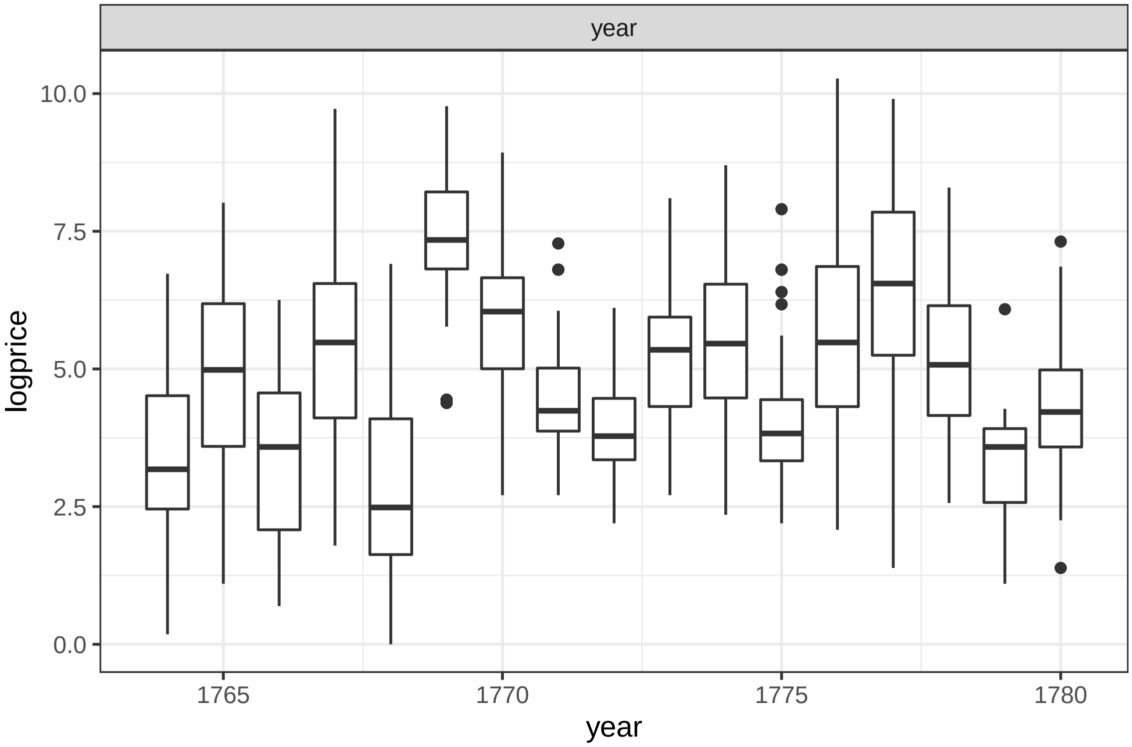

Figure 3: Boxplots between logprice and year

- 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.
- 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.

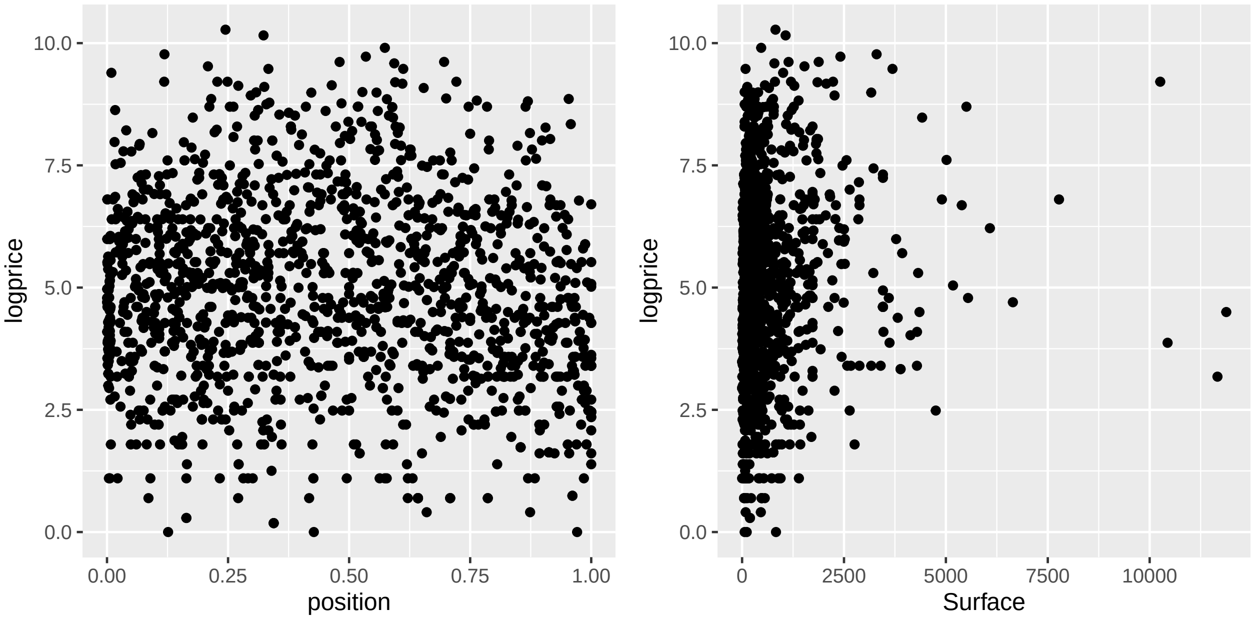

Figure 4: Scatterplots between logprice and position as well as log(Surface)

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.

Development and Assessment of Initial Model

Initial Model

Before we start building a model, we first first convert a few variable to factor(the binary variables and year), and then we change the category "EB" to "E" for the variable `type_intermed`, because we have only one observation with that category. We change it to "E" because the primary is an expert.

Model Selection

From the EDA, we first considered all of the variables deemed important, along with interactions we thought would be important. We thought certain binary features may/may not be as important depending on the type of buyer. Certain interactions could relate to all people involved (dealer, endbuyer, and potential intermediaries), while other interactions really only seem relevant to the buyers (endbuyer and intermediary only).

```
##
## Call:
## lm(formula = logprice ~ (dealer + endbuyer + type_intermed) *
##     (diff_origin + discauth + lrgfont + engraved + prevcoll +
##         paired) + (endbuyer + type_intermed) * (figures + finished +
##     portrait + still_life + peasant + othgenre + lands_sc + lands_elem) +
##     year + origin_author, data = paintings_train)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.2189 -0.7127  0.0000  0.6750  3.7162 
## 
## Coefficients: (15 not defined because of singularities)
##                             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                  2.04647    0.48589   4.212 2.70e-05 ***
## dealerL                      1.77206    0.29620   5.983 2.82e-09 ***
## dealerP                      0.92858    0.31193   2.977 0.002964 ** 
## dealerR                      1.45091    0.17845   8.131 9.64e-16 ***
## endbuyerB                    1.49666    0.74043   2.021 0.043443 *  
## endbuyerC                    0.72551    0.21465   3.380 0.000746 ***
## endbuyerD                    0.90084    0.18148   4.964 7.80e-07 ***
## endbuyerE                    0.01062    0.24470   0.043 0.965388    
## endbuyerU                    0.45369    0.22665   2.002 0.045517 *  
## type_intermedB               2.60286    2.05379   1.267 0.205254    
## type_intermedD               1.15389    0.33885   3.405 0.000680 ***
## type_intermedE               1.10603    0.49576   2.231 0.025849 *  
## diff_origin1                 0.60442    0.29604   2.042 0.041378 *  
## discauth1                    0.93831    0.39165   2.396 0.016720 *  
## lrgfont1                     0.59199    0.61932   0.956 0.339314    
## engraved1                   -0.03721    0.68174  -0.055 0.956481    
## prevcoll1                    0.30909    0.43010   0.719 0.472490    
## paired1                     -0.05296    0.20669  -0.256 0.797799    
## figures1                    -0.14743    0.24673  -0.598 0.550260    
## finished1                    0.23133    0.24489   0.945 0.345008    
## portrait1                   -0.58437    0.27519  -2.124 0.033891 *  
## still_life1                 -0.89116    0.25614  -3.479 0.000519 ***
## peasant1                     0.08314    0.27931   0.298 0.766012    
## othgenre1                    0.36295    0.25699   1.412 0.158086    
## lands_sc1                   -0.22755    0.22807  -0.998 0.318600    
## lands_elem1                  0.04423    0.14115   0.313 0.754036    
## year1765                     1.48695    0.21286   6.985 4.46e-12 ***
## year1766                    -0.17415    0.21927  -0.794 0.427213    
## year1767                     1.37650    0.14678   9.378  < 2e-16 ***
## year1768                     0.11930    0.17167   0.695 0.487222    
## year1769                     1.89681    0.27685   6.852 1.11e-11 ***
## year1770                     1.06834    0.22314   4.788 1.87e-06 ***
## year1771                     1.14616    0.17957   6.383 2.39e-10 ***
## year1772                     0.70061    0.24359   2.876 0.004089 ** 
## year1773                     1.10247    0.20403   5.404 7.73e-08 ***
## year1774                     1.68335    0.17573   9.579  < 2e-16 ***
## year1775                     1.40739    0.28721   4.900 1.07e-06 ***
## year1776                     1.56497    0.14528  10.772  < 2e-16 ***
## year1777                     2.30167    0.15232  15.111  < 2e-16 ***
## year1778                     1.35070    0.27660   4.883 1.17e-06 ***
## year1779                     0.70704    0.33384   2.118 0.034367 *  
## year1780                     1.07041    0.30120   3.554 0.000393 ***
## origin_authorD/FL           -0.10260    0.44827  -0.229 0.818998    
## origin_authorF              -0.37828    0.44870  -0.843 0.399357    
## origin_authorG              -0.39689    0.49259  -0.806 0.420556    
## origin_authorI              -0.62452    0.45546  -1.371 0.170544    
## origin_authorS              -0.19848    0.58382  -0.340 0.733932    
## origin_authorX              -1.19236    0.44849  -2.659 0.007940 ** 
## dealerL:diff_origin1        -0.07172    0.35185  -0.204 0.838509
## dealerP:diff_origin1        -0.79248    0.34557  -2.293 0.021987 *
## dealerR:diff_origin1        -0.95380    0.25165  -3.790 0.000157 ***
## dealerL:discauth1            0.08105    0.89295   0.091 0.927692
## dealerP:discauth1           -3.66300    0.96260  -3.805 0.000148 ***
## dealerR:discauth1           -1.38682    0.32212  -4.305 1.79e-05 ***
## dealerL:lrgfont1                  NA         NA      NA       NA
## dealerP:lrgfont1                  NA         NA      NA       NA
## dealerR:lrgfont1                  NA         NA      NA       NA
## dealerL:engraved1            0.27194    0.96570   0.282 0.778298
## dealerP:engraved1                 NA         NA      NA       NA
## dealerR:engraved1           -0.29677    0.51411  -0.577 0.563867
## dealerL:prevcoll1            0.84830    0.62513   1.357 0.175015
## dealerP:prevcoll1            2.29383    1.04712   2.191 0.028652 *
## dealerR:prevcoll1            0.83049    0.70194   1.183 0.236963
## dealerL:paired1             -1.06826    0.25066  -4.262 2.17e-05 ***
## dealerP:paired1             -0.65037    0.31815  -2.044 0.041129 *
## dealerR:paired1             -0.28611    0.20364  -1.405 0.160253
## endbuyerB:diff_origin1      -1.13587    1.34587  -0.844 0.398838
## endbuyerC:diff_origin1      -0.70986    0.29664  -2.393 0.016850 *
## endbuyerD:diff_origin1      -0.30574    0.24435  -1.251 0.211070
## endbuyerE:diff_origin1       0.15089    0.34166   0.442 0.658821
## endbuyerU:diff_origin1      -0.24703    0.29268  -0.844 0.398813
## endbuyerB:discauth1          0.32433    1.96516   0.165 0.868938
## endbuyerC:discauth1         -0.75152    0.63199  -1.189 0.234598
## endbuyerD:discauth1          0.17237    0.45323   0.380 0.703775
## endbuyerE:discauth1          0.54296    0.52704   1.030 0.303104
## endbuyerU:discauth1         -0.24578    0.50933  -0.483 0.629485
## endbuyerB:lrgfont1          -0.71525    1.05622  -0.677 0.498407
## endbuyerC:lrgfont1           0.35206    0.65768   0.535 0.592528
## endbuyerD:lrgfont1           0.20966    0.64471   0.325 0.745076
## endbuyerE:lrgfont1           0.34350    0.81905   0.419 0.674997
## endbuyerU:lrgfont1           0.67835    0.78630   0.863 0.388449
## endbuyerB:engraved1               NA         NA      NA       NA
## endbuyerC:engraved1          1.24844    0.66088   1.889 0.059101 .
## endbuyerD:engraved1          1.38997    0.63459   2.190 0.028671 *
## endbuyerE:engraved1          1.07172    0.68544   1.564 0.118160
## endbuyerU:engraved1          1.20270    0.69394   1.733 0.083298 .
## endbuyerB:prevcoll1               NA         NA      NA       NA
## endbuyerC:prevcoll1          0.02981    0.72598   0.041 0.967258
## endbuyerD:prevcoll1         -0.19036    0.67839  -0.281 0.779053
## endbuyerE:prevcoll1          0.04033    0.73632   0.055 0.956330
## endbuyerU:prevcoll1         -0.28569    0.93389  -0.306 0.759721
## endbuyerB:paired1            0.31761    1.11932   0.284 0.776643
## endbuyerC:paired1            0.03717    0.25834   0.144 0.885626
## endbuyerD:paired1            0.04357    0.20474   0.213 0.831508
## endbuyerE:paired1            0.05998    0.29309   0.205 0.837886
## endbuyerU:paired1            0.13860    0.25619   0.541 0.588585
## type_intermedB:diff_origin1  2.14684    1.41756   1.514 0.130145
## type_intermedD:diff_origin1  0.43869    0.52916   0.829 0.407231
## type_intermedE:diff_origin1 -0.71947    0.79980  -0.900 0.368514
## type_intermedB:discauth1     2.81689    1.43233   1.967 0.049431 *
## type_intermedD:discauth1     1.91897    0.71125   2.698 0.007063 **
## type_intermedE:discauth1     1.56265    0.89342   1.749 0.080509 .
## type_intermedB:lrgfont1     -1.49000    1.91183  -0.779 0.435908    
## type_intermedD:lrgfont1     -0.33077    0.36411  -0.908 0.363812    
## type_intermedE:lrgfont1      0.63107    0.54503   1.158 0.247132    
## type_intermedB:engraved1          NA         NA      NA       NA    
## type_intermedD:engraved1     0.04350    0.47025   0.093 0.926306    
## type_intermedE:engraved1    -0.97843    0.75481  -1.296 0.195112    
## type_intermedB:prevcoll1          NA         NA      NA       NA    
## type_intermedD:prevcoll1    -0.62335    0.50022  -1.246 0.212930    
## type_intermedE:prevcoll1     0.25501    0.76231   0.335 0.738035    
## type_intermedB:paired1      -0.85766    2.36271  -0.363 0.716662    
## type_intermedD:paired1      -0.40503    0.31413  -1.289 0.197485    
## type_intermedE:paired1      -1.29920    0.52200  -2.489 0.012935 *  
## endbuyerB:figures1          -0.78513    1.54511  -0.508 0.611441    
## endbuyerC:figures1           0.29047    0.40080   0.725 0.468747    
## endbuyerD:figures1           0.71658    0.34044   2.105 0.035490 *  
## endbuyerE:figures1           0.40549    0.52443   0.773 0.439536    
## endbuyerU:figures1           0.74829    0.57917   1.292 0.196583    
## endbuyerB:finished1         -0.14622    1.13258  -0.129 0.897297    
## endbuyerC:finished1          0.71532    0.33585   2.130 0.033365 *  
## endbuyerD:finished1          0.17349    0.28509   0.609 0.542930    
## endbuyerE:finished1          0.75901    0.37389   2.030 0.042552 *  
## endbuyerU:finished1         -0.14694    0.36782  -0.399 0.689598    
## endbuyerB:portrait1               NA         NA      NA       NA    
## endbuyerC:portrait1          0.69965    0.63494   1.102 0.270699    
## endbuyerD:portrait1          0.27109    0.42456   0.639 0.523244    
## endbuyerE:portrait1          0.51880    0.50266   1.032 0.302207    
## endbuyerU:portrait1         -1.15604    0.65096  -1.776 0.075979 .  
## endbuyerB:still_life1             NA         NA      NA       NA    
## endbuyerC:still_life1        1.17906    0.58474   2.016 0.043959 *  
## endbuyerD:still_life1        0.57212    0.39045   1.465 0.143074    
## endbuyerE:still_life1        0.93657    0.91092   1.028 0.304062    
## endbuyerU:still_life1        0.18982    0.54109   0.351 0.725781    
## endbuyerB:peasant1          -0.26390    1.74951  -0.151 0.880125    
## endbuyerC:peasant1           0.12749    0.48153   0.265 0.791231    
## endbuyerD:peasant1           0.03291    0.35597   0.092 0.926357    
## endbuyerE:peasant1           0.92791    0.56888   1.631 0.103101    
## endbuyerU:peasant1          -0.32849    0.50476  -0.651 0.515295    
## endbuyerB:othgenre1         -0.35065    2.04940  -0.171 0.864172    
## endbuyerC:othgenre1         -0.04030    0.36930  -0.109 0.913113    
## endbuyerD:othgenre1          0.01638    0.32359   0.051 0.959637    
## endbuyerE:othgenre1          0.71091    0.43761   1.625 0.104502    
## endbuyerU:othgenre1         -0.04900    0.43590  -0.112 0.910512    
## endbuyerB:lands_sc1          0.04327    1.34236   0.032 0.974291    
## endbuyerC:lands_sc1         -0.24515    0.39826  -0.616 0.538295    
## endbuyerD:lands_sc1         -0.53712    0.30208  -1.778 0.075620 .  
## endbuyerE:lands_sc1          0.73782    0.46094   1.601 0.109685    
## endbuyerU:lands_sc1         -0.34841    0.36783  -0.947 0.343704    
## endbuyerB:lands_elem1        0.52673    1.18765   0.444 0.657471    
## endbuyerC:lands_elem1        0.35718    0.24071   1.484 0.138073    
## endbuyerD:lands_elem1        0.07558    0.18545   0.408 0.683677    
## endbuyerE:lands_elem1        0.25942    0.31012   0.837 0.403014    
## endbuyerU:lands_elem1        0.04149    0.25762   0.161 0.872089    
## type_intermedB:figures1           NA         NA      NA       NA    
## type_intermedD:figures1     -0.25732    0.54111  -0.476 0.634485    
## type_intermedE:figures1    -0.76956    1.25090  -0.615 0.538526    
## type_intermedB:finished1   -0.79254    1.13064  -0.701 0.483446    
## type_intermedD:finished1   -0.33800    0.40029  -0.844 0.398608    
## type_intermedE:finished1   -0.84205    0.58691  -1.435 0.151598    
## type_intermedB:portrait1   -4.10944    2.13840  -1.922 0.054851 .  
## type_intermedD:portrait1   -0.80683    1.28776  -0.627 0.531067    
## type_intermedE:portrait1   -1.85848    0.95526  -1.946 0.051920 .  
## type_intermedB:still_life1 -2.84141    1.68985  -1.681 0.092906 .  
## type_intermedD:still_life1  0.99890    0.80510   1.241 0.214926    
## type_intermedE:still_life1 -0.41362    1.09214  -0.379 0.704954    
## type_intermedB:peasant1          NA         NA      NA       NA    
## type_intermedD:peasant1     0.92538    0.83703   1.106 0.269116    
## type_intermedE:peasant1     0.84666    1.51414   0.559 0.576141    
## type_intermedB:othgenre1         NA         NA      NA       NA    
## type_intermedD:othgenre1   -0.55244    0.44695  -1.236 0.216660    
## type_intermedE:othgenre1   -0.14514    0.76886  -0.189 0.850294    
## type_intermedB:lands_sc1         NA         NA      NA       NA    
## type_intermedD:lands_sc1         NA         NA      NA       NA    
## type_intermedE:lands_sc1   -1.13818    0.99054  -1.149 0.250740    
## type_intermedB:lands_elem1 -2.41348    2.35714  -1.024 0.306066    
## type_intermedD:lands_elem1 -0.06854    0.34824  -0.197 0.844000    
## type_intermedE:lands_elem1 -0.13438    0.51035  -0.263 0.792357    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.103 on 1337 degrees of freedom
## Multiple R-squared:  0.705,  Adjusted R-squared:  0.6692 
## F-statistic: 19.72 on 162 and 1337 DF,  p-value: < 2.2e-16
```

However, this is far too many variables; many of these are likely noise and aren't truly associated with price. We can use stepwise variable selection with BIC to select the most important variables. There are several methods for reducing the number of variables, typically comparing AIC, BIC, and adjusted-R squared. We are opting for BIC because it penalizes more harshly than the other 2, and we are trying to achieve a parsimonious model for this part of our analysis. Additionally, there are also several interactions with NA components, and stepwise BIC may remove these terms. The reason for these NAs might be that there is not enough data for those specific interactions or there might be an issue of correlation present between some of the predictors.

The model has an adjusted R-squared value of 0.669 which is acceptable, but even if it could explain a lot of the variation in the training data, we cannot say it will perform better on the test data. Also, we wouldn't want to use this model for interpretation or prediction because of the presence of NAs.

While we can use the adjusted R-squared to compare between models, the most important metric is how well our model will perform on new data. A typical way to measure this is RMSE, a function of the residuals which we define below. We first see how well our model performs via the training RMSE and the coverage. Our model here should be adequate, as we are using this data to fit the model. But we also need to compute the test RMSE and coverage. If there is a drastic drop between the training and the test RMSE, we can conclude that our model is likely overfitting. We need to reduce the number of variables to get the model to generalize more to new data.

By itself, our training RMSE cannot really tell us much; we need to compare it to the test RMSE first. However, our coverage is very good. It should be ~95%, and it more or less is exactly that.

However, our testing data is less good. Currently, we cannot check the test RMSE, but we can see the test coverage, which was around 65%. This is evidence of overfitting, so we would like to improve our model by removing even more variables that are not explaining the variation in price enough. We removed `year` because we found that certain dealers were only dealing in certain years, so much of the explanatory power of year was already captured by `dealer`. With `origin_author`, we simply removed it due to a lack of significance. After we removed these two variables, we ran the model on the test data and found it's still likely overfitting.

We think after looking at the data, a lot of the paintings' features look correlated with the dealers of the painting, especially `lrgfont`. So, maybe removing `dealer` from the model can help. We observed that the coverage in the test data improved, but the coverage was still much lower than on the training data.

We finally removed `type_intermed` from our model and our coverage on the test data improved again. While the coverage was still much lower than on the training data, removing any more variables may bring too much bias in our model along with a low R-squared value. So, we started with a model with the `endbuyer` and its interactions with the binary variables, and then used stepwise selection with BIC to get to the final model. We find that we end up with a model with 12 predictors and no interactions.

Final Model Discussion

Our final model is as follows:

```
##
## Call:
## lm(formula = logprice ~ endbuyer + diff_origin + lrgfont + engraved +
##     prevcoll + finished + paired + portrait + still_life + othgenre +
##     lands_sc + lands_elem, data = paintings_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8572 -0.9177  0.0002  0.9094  5.4727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.13577    0.09703  42.624  < 2e-16 ***
## endbuyerB    1.47064    0.38839   3.786 0.000159 ***
## endbuyerC    1.48855    0.11598  12.835  < 2e-16 ***
## endbuyerD    1.09852    0.10143  10.830  < 2e-16 ***
## endbuyerE    0.24022    0.14783   1.625 0.104389
## endbuyerU    0.31953    0.13298   2.403 0.016394 *
## diff_origin1 -0.87176    0.08944  -9.747  < 2e-16 ***
## lrgfont1     1.51577    0.13554  11.183  < 2e-16 ***
## engraved1    0.48848    0.17104   2.856 0.004350 **
## prevcoll1    1.24165    0.16961   7.321 4.03e-13 ***
## finished1    0.43649    0.10523   4.148 3.54e-05 ***
## paired1     -0.44729    0.07786  -5.744 1.12e-08 ***
## portrait1   -0.78699    0.20046  -3.926 9.04e-05 ***
## still_life1 -0.64621    0.19456  -3.321 0.000918 ***
## othgenre1    0.58817    0.13227   4.447 9.37e-06 ***
## lands_sc1   -0.51886    0.14210  -3.651 0.000270 ***
## lands_elem1  0.37305    0.08525   4.376 1.29e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.416 on 1483 degrees of freedom
## Multiple R-squared:  0.4608, Adjusted R-squared:  0.455
## F-statistic: 79.22 on 16 and 1483 DF,  p-value: < 2.2e-16
```

Assumptions of linear regression

There are several assumptions we need to investigate to determine whether our model is valid or not. We can see most of these with the following plots

Discussion:

1) From the QQ-plot, the residuals seem to be normally distributed.
2) From the Residuals vs Fitted and Scale-Location plot, it seems that the residual can be explained well by our model.
3) From the Residuals vs Leverage plot, it seems there are no outliers and high leverage points.

Table of variables

We need to exponentiate the point estimates and the confidence intervals to convert to the non-log scale of price (in livres).

Table 1: Table of coefficients and confidence intervals

	Fit	Lower	Upper
(Intercept)	62.538	51.699	75.648
endbuyerB	4.352	2.032	9.323
endbuyerC	4.431	3.529	5.563
endbuyerD	3.000	2.459	3.660
endbuyerE	1.272	0.951	1.699
endbuyerU	1.376	1.060	1.787
diff_origin1	0.418	0.351	0.498
lrgfont1	4.553	3.490	5.940
engraved1	1.630	1.165	2.280
prevcoll1	3.461	2.482	4.828
finished1	1.547	1.259	1.902
paired1	0.639	0.549	0.745
portrait1	0.455	0.307	0.675
still_life1	0.524	0.358	0.768
othgenre1	1.801	1.389	2.334
lands_sc1	0.595	0.450	0.787
lands_elem1	1.452	1.229	1.716

Summary and Conclusions

The median price for the baseline category corresponds to exp(β_0), which is 62.538 (with a 95% confidence interval of (51.699, 75.648)). This corresponds to the median price in livres for a painting whose buyer has no information (even name), and whose is featureless for the binary variables listed above.

Important variables

To find out important predictors in our model, we typically do not want to directly compare coefficient values, as these change with the scale of the predictor. However, because all of our variables are categorical (exist outside of a scale), we can in fact compare their coefficients to test for importance. Using this as our metric, the most important variables are:

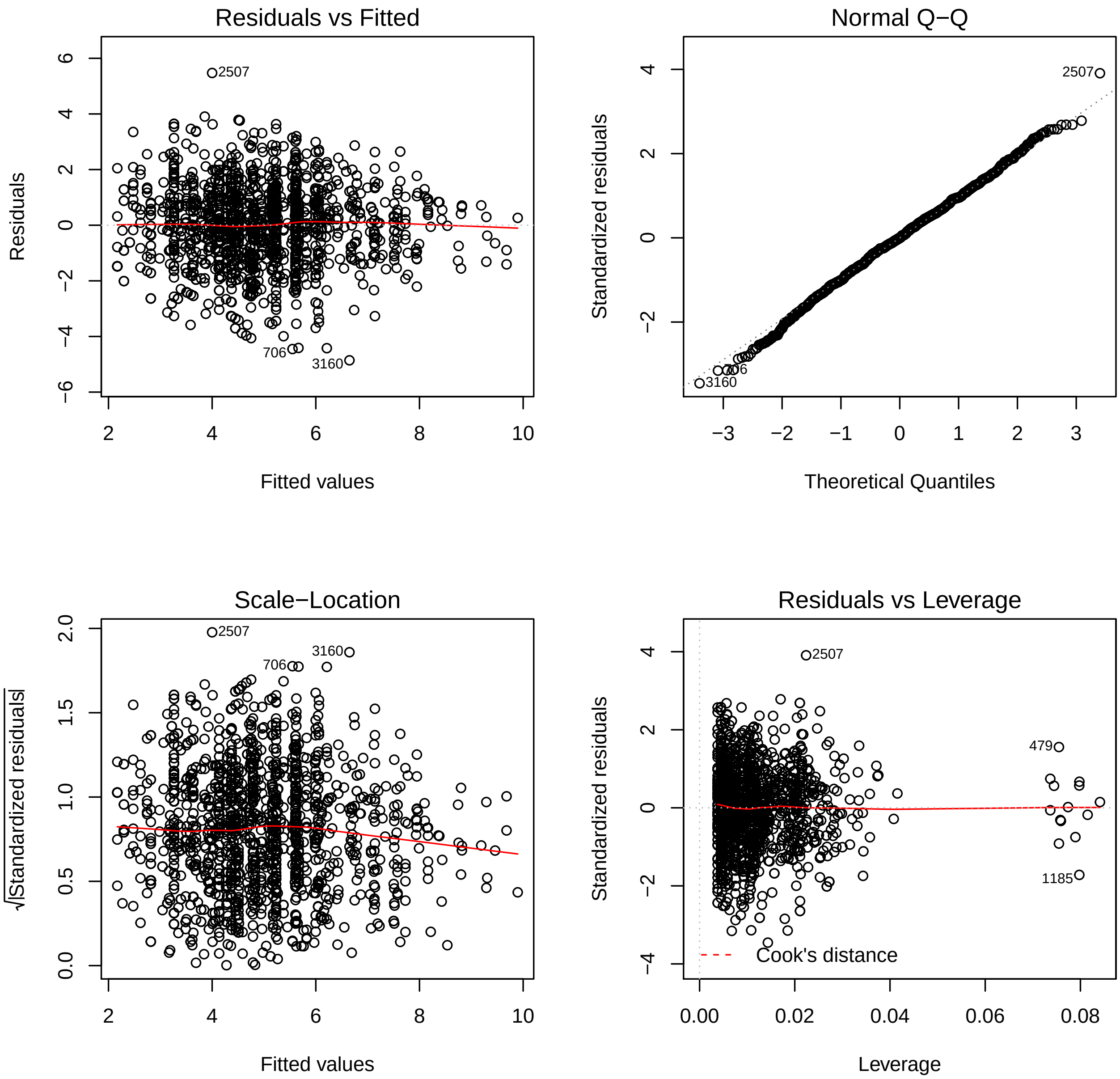

Figure 5: Diagnostic plots of the final model

- `endbuyer`: type of endbuyer
- `diff_origin`: is the painter's nationality different than listed in the catalogue?
- `lrgfont`: does the dealer devote an additional paragraph to the painting (listed in larger font)?
- `prevcoll`: is the previous owner mentioned?

Note that our final model does not contain any interactions. This is a consequence of using the stepwise procedure to reduce model complexity.

Variable interpretation

We can interpret the coefficients of the most important variables listed above as follows, keeping other predictors fixed:

- β_{diff_origin}: We expect the median price in livers to decrease by 58%(with a confidence interval of (50%,65%)) when the nationality listed in the catalogue is different from the painter's.
- $\beta_{lrgfont}$: We expect the median price in livers to to increase by 355%(with a confidence interval of (249%,494%)) when the dealer devotes an additional paragraph to the painting in large font.
- $\beta_{prevcoll}$: We expect the median price in livers to to increase by 246%(with a confidence interval of (148%,383%)) if the previous owner is mentioned.
- Coefficients for endbuyer:
  1. $\beta_{endbuyerB}$: We expect the median price in livers to increase by 335.2% (with a confidence interval of (103.2%, 832.3%)) when the endbuyer is also a buyer.
  2. $\beta_{endbuyerC}$: We expect the median price in livers to increase by 343.1% (with a confidence interval of (252.9%, 456.3%)) when the endbuyer is a collector.
  3. $\beta_{endbuyerD}$: We expect the median price in livers to increase by 200% (with a confidence interval of (145.9%, 266%)) when the endbuyer is a dealer.
  4. $\beta_{endbuyerE}$: We expect the median price in livers to increase by 27.2% (with a confidence interval of (-4.9%, 69.9%)) when the endbuyer is an expert.
  5. $\beta_{endbuyerU}$: We expect the median price in livers to increase by 37.6% (with a confidence interval of (6%, 78.7%)) when the endbuyer is unknown.

Painting Recommendations

To find the most valuable paintings and to make a recommendation to the art historian, we look at the results of our final model.
Paintings where the dealer had a description in large font and the previous owner was mentioned, seem to be much more valuable than those which didn't. The painting was more valuable if the dealer had mentioned engravings done after the painting. Apart from these, the painting was more valuable if the painting was noted for its highly polished finishing, had a description that mentioned a genre scene other than a peasant scene and had landscape elements mentioned in its description.

Since, the data is historic, we have details of the buyers of the paintings as well. It seems that if the end buyer was a normal buyer or a collector, then the painting was likely to be more valuable(more than 4 times), it was more valuable if the buyer was a dealer as well(about 3 times). The painting was likely to be less valuable if the end buyer was an expert or unknown as compared to a buyer, collector or dealer but more valuable if there was absolutely no information about the buyer, not even the name.

Finally, paintings which had different origins in the catalogue when compared to the origin of the artist, was suggested or sold as a pairing with another painting, was described as a portrait, had a description which mentioned still life elements or was described as a plain landscape were valued less.

Findings and potential limitations

Our model finds that with our current chosen variables, not adding interaction terms seems to be better to do the prediction than adding interaction terms. As for the limitations of our model, intuitively, the price of the painting may be related to the sale of the year (because if economy is better in some year, people might want to spend more money on a painting since they might have more income to spend) and dealer (because just like buying a car, different dealers have different prices even for the same products), but these are the variables that we decided not to include in our final model. Therefore, lacking these two variables may limit our model to capture as much information as possible to do the prediction. Besides, we didn't use some predictors like surface, because it has some missing values and this variable may have the potential to capture some information that other variables cannot capture. Therefore, we may miss some important information from those variables that we chose not to include. In addition, using a linear model just can't tell us which actual set of predictors are the best since it cannot account for the collinearity in the data set if the collinearity is literally presented in the predictors that are included in our final model.