# project

*Team Omega*

*December 3, 2017*

## 1. Introduction:

## Summary of Problem:

We are given a dataset which provides information about auctions of paintings which were sold between years 1764 and 1780. The data tells us about different attributes of paintings, information about auction, information about the artist, information about the buyer and the price at which it was sold. There are total 59 columns (variables) in data including 'price' and 'log(price)'. Our task is to find out the relation between variables and the price so that we can predict the price of a paintings. We should also interpret the results as which are the most/least influential factor in determining the price of the painting.

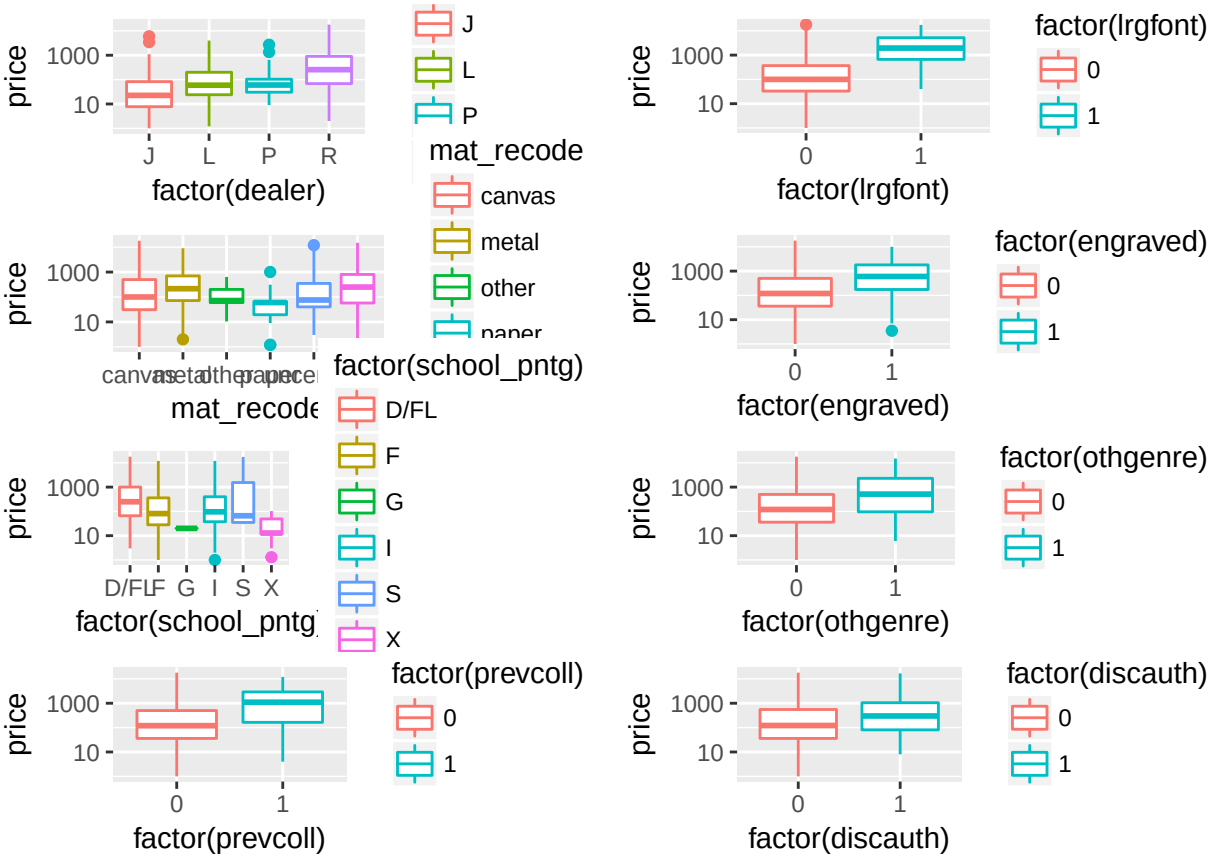
Objectives:

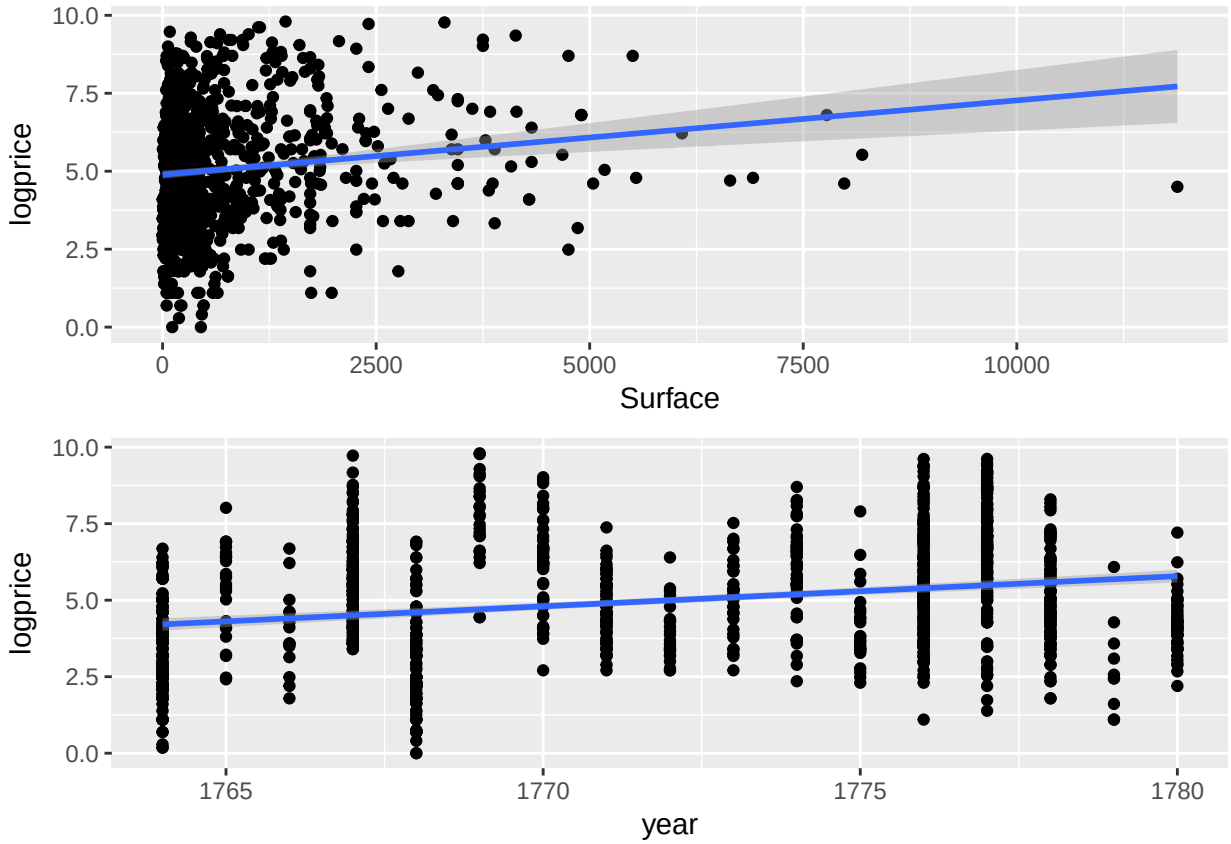
1) Exploratory data analysis
2) Find out interaction
3) Create a model with 10-20 most influential factors and interpret results
4) Test the model with test data and interpret the results

## 2. Exploratory Data Analysis

### Some Graphs and analysis

The following is a group of boxplots that show the effect of different variables on price. Based on these boxplots, we observe that the mean of price differ between layers of categorical variables; the discrepancies are later on substantiated by hypothesis testing to be statistically significant. As a preliminary analysis, we decide to put them into our raw model before BIC analysis. It turns out that all but mat_recode are included in our final model. Even though mat_recode is not significant on its own, it proves to be a very interesting variable in interaction term.






**Interactions**

(1). Surface:mat_recode

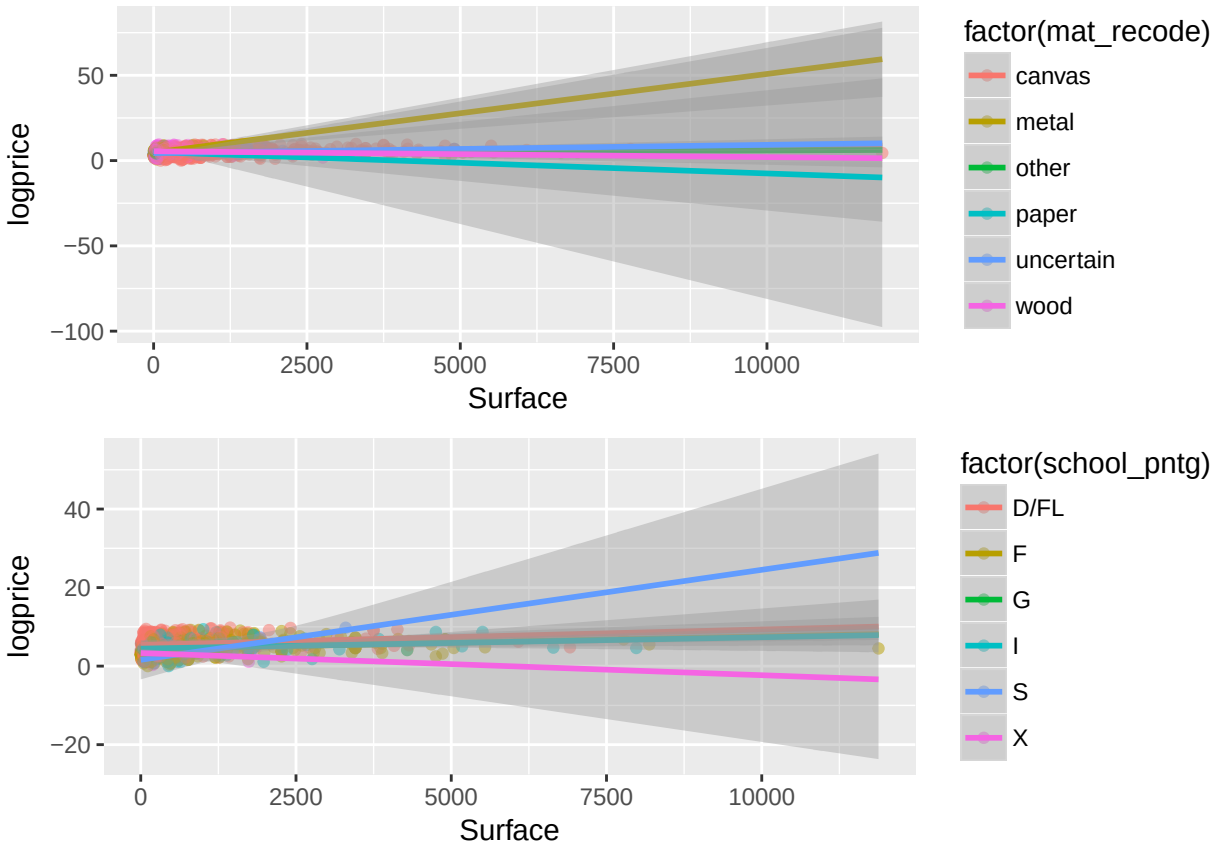
Surface area and mat_recode (recoding of material) should have some interaction as different material is used fo paintings for different sizes of paintings. This is further supported as we observe significantly different slopes from layer to layer.

(2). school_pntg:Surface

School of painting and surface area of painting will definitely have some interaction as different style of paintings will use different figures and landscapes in them which will have diffeent sizes. For example, a portrait will usually have smaller surface area than a landscape which includes mountains and rivers.

(3). winningbiddertype:prevcoll

One of the motivating factors of purchasing a painting could be the name of previous owner of that painting. If bidders know that the previous owner of the painting was a well-known person, then they may be more willing to pay for a higher price. So we figure that 'winningbiddertype' and 'prevcoll' may have some interaction.


logprice
50
0
−50
−100
0
2500
5000
7500
10000
Surface
factor(mat_recode)
canvas
metal
other
paper
uncertain
wood
logprice
40
20
0
−20
0
2500
5000
7500
10000
Surface
factor(school_pntg)
D/FL
F
G
I
S
X

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

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

```
##
## Call:
## lm(formula = log(price) ~ dealer + year + school_pntg + diff_origin +
##     artistliving + winningbiddertype + Surface + engraved + prevcoll +
##     paired + finished + lrgfont + othgenre + discauth + winningbiddertype:prevcoll +
##     Surface:mat_recode, data = newTrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5612 -0.7434  0.0007  0.6965  3.5332
##
## Coefficients: (6 not defined because of singularities)
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -1.510e+02  1.569e+01  -9.620  < 2e-16 ***
## dealerL                        1.575e+00  1.527e-01  10.316  < 2e-16 ***
## dealerP                        4.358e-01  1.964e-01   2.219 0.026696 *
## dealerR                        1.779e+00  1.291e-01  13.783  < 2e-16 ***
## year                           8.684e-02  8.851e-03   9.811  < 2e-16 ***
## school_pntgF                  -6.885e-01  8.839e-02  -7.790 1.63e-14 ***
## school_pntgI                  -7.039e-01  1.128e-01  -6.241 6.33e-10 ***
## school_pntgOTHER              -8.835e-01  5.266e-01  -1.678 0.093685 .
## school_pntgX                  -1.095e+00  3.355e-01  -3.263 0.001138 **
## diff_origin                   -6.646e-01  9.649e-02  -6.888 9.80e-12 ***
## artistliving                   3.293e-01  1.250e-01   2.634 0.008562 **
## winningbiddertypeB             1.140e+00  3.561e-01   3.200 0.001417 **
## winningbiddertypeBC            1.596e+00  3.877e-01   4.117 4.14e-05 ***
## winningbiddertypeC             8.490e-01  1.556e-01   5.457 6.07e-08 ***
## winningbiddertypeD             8.087e-01  1.297e-01   6.237 6.49e-10 ***
## winningbiddertypeDB            1.792e+00  8.424e-01   2.128 0.033613 *
## winningbiddertypeDC            2.169e+00  1.875e-01  11.566  < 2e-16 ***
## winningbiddertypeDD            1.791e+00  4.966e-01   3.606 0.000326 ***
## winningbiddertypeE             3.668e-01  1.653e-01   2.219 0.026729 *
## winningbiddertypeEB            1.891e-01  8.295e-01   0.228 0.819760
## winningbiddertypeOTHER         2.256e+00  8.429e-01   2.677 0.007551 **
## winningbiddertypeEC            1.572e+00  2.474e-01   6.353 3.17e-10 ***
## winningbiddertypeU             4.841e-01  1.485e-01   3.260 0.001149 **
## Surface                        3.740e-04  3.585e-05  10.432  < 2e-16 ***
## engraved                       5.297e-01  1.669e-01   3.173 0.001553 **
## prevcoll                       7.653e-01  3.825e-01   2.001 0.045688 *
## paired                        -2.957e-01  7.797e-02  -3.793 0.000158 ***
## finished                       6.513e-01  1.058e-01   6.154 1.08e-09 ***
## lrgfont                        9.040e-01  1.318e-01   6.859 1.19e-11 ***
## othgenre                       4.688e-01  1.242e-01   3.776 0.000169 ***
## discauth                       5.014e-01  1.538e-01   3.261 0.001147 **
## winningbiddertypeB:prevcoll           NA         NA      NA       NA
## winningbiddertypeBC:prevcoll          NA         NA      NA       NA
## winningbiddertypeC:prevcoll    1.056e-01  5.751e-01   0.184 0.854349
## winningbiddertypeD:prevcoll   -3.896e-02  4.515e-01  -0.086 0.931249
## winningbiddertypeDB:prevcoll          NA         NA      NA       NA
```

```
## winningbiddertypeDC:prevcoll    -1.049e+00  5.549e-01  -1.891 0.058965 .
## winningbiddertypeDD:prevcoll            NA         NA      NA       NA
## winningbiddertypeE:prevcoll      2.888e-01  6.242e-01   0.463 0.643646
## winningbiddertypeEB:prevcoll            NA         NA      NA       NA
## winningbiddertypeOTHER:prevcoll         NA         NA      NA       NA
## winningbiddertypeEC:prevcoll     5.645e-02  8.080e-01   0.070 0.944314
## winningbiddertypeU:prevcoll     -1.896e+00  1.228e+00  -1.544 0.122803
## Surface:mat_recodemetal          1.349e-03  5.717e-04   2.359 0.018509 *
## Surface:mat_recodeother          4.348e-04  1.391e-03   0.313 0.754608
## Surface:mat_recodepaper          1.643e-04  1.563e-03   0.105 0.916310
## Surface:mat_recodeuncertain     -1.940e-05  1.319e-04  -0.147 0.883069
## Surface:mat_recodewood           1.526e-05  1.338e-04   0.114 0.909207
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.156 on 1034 degrees of freedom
## Multiple R-squared:  0.6422, Adjusted R-squared:  0.628
## F-statistic: 45.26 on 41 and 1034 DF,  p-value: < 2.2e-16
```

As we observe from the summary, the first part we notice is that there are several NA values in some layers of our interaction terms, because the training dataset may not have data to back up the layer, thus generating NA values. After that, we can interpret the coefficients of categorical variables in a sense that when the variable is changed from 0 (reference level) to 1, the value of log(price) will change by an amount of $\beta$. For example, the variable "lrgfont" has a coefficient 9.040e-01. This means that if we change "lrgfont" from 0 to 1 (ie. if the dealer devotes an additional paragraph), then on average, the log(price) will increase by 9.040e-01, which in turn translates to an increase of $e^{9.040e-01} = 2.479359$ livres. For continuous variable, such as "Surface", the coefficient represent that for every 1 unit increase in predictor, we expect the price to change by an amount of $\beta$. For example, the variable "Surface" has a coefficient of 3.740e-04. That is, for every 1-unit increase in "Surface", we will have on average an increase in price by $e^{3.740e-04} = 1$ livres.

### Model selection

Since we are given a dataset with 59 variables (including 'price' and 'log(pirce)'), a model that includes all of them will be too complicated and run into the risk of overfitting. Therefore, we decide to do model selection by removing some variables that may not help our prediction. We carry out the model selection by four steps: recoding, general analysis, BIC, and interaction.

To begin with, we decide to recode some variables to better represent their respective features. Besides the recode of 'shape_recode' provided by Professor, we also generate some recode variables: 'mat_recode', 'fig_mention', 'artist_living_notliving', 'history_nohistory', 'mytho_nomytho', 'finished_nofinished', and 'LF'. In particular, in 'mat_recode', we recode "a" (silver), "bc" (wood and copper), "c" (copper),"br" (bronze frames) into "metal"; "al" (alabaster), "ar" (slate), "m" (marble) into "stone"; "co" (cloth), "bt" (canvas), "t" (canvas),"h" (oil technique),"ta" (canvas) into "canvas"; "p" (paper), "ca" (cardboard) into "paper"; "b" (wood) into "wood"; "o" (other), "e" (wax),"v" (glass),"mi" (miniature technique),"pa" (pastel),"g" (grissaille technique) into "other"; the rest into "uncertain". For the other recode variables, we use them as indicators of corresponding features (0 if the feature does not exist, and 1 if it does).

After recoding, we remove some variables which we believe are not contributing to our model or are repetitive. After careful analysis, we decide to remove "sale", "lot", "position", "logprice", "subject", "authorstyle", "authorstandard", "author", "winningbidder", "Interm", "Height_in", "Width_in", "Surface_Rect", "Diam_in", "Surface_Rnd", "material", "mat", "lands_sc", "lands_elem", "lands_figs", and "lands_ment". In particular, we remove "sale" because it is a combination of dealer and year, both of which we have included in our raw model. We remove "lot" and "position" because they are identifiers of paintings, which we believe are not helpful in linear modeling. We remove "logprice" because we can just take log(price). We remove "subject" because it is descriptive, and even though we may be able to use some Natural Language Processing

(NLP) knowledge to understand this feature better, this is out of the scope of our current task. We remove "authorstyle" because this variable contains too many n/a, which itself has no meaning (sometimes, a value is missing for some particular reason, but in this case we do not think so), and we have very few method to interpret these n/a values. We remove both "authorstandard" and "author" because they contain too many layers; while they could be useful predictors if studied carefully (paintings associated with certain authors may be more expensive than others), we decide to drop them as we do not have sufficient information about each of them. We remove "winningbidder" because its information is conveyed by "winningbiddertype". We remove "Interm" for the same reason as previous one. We remove "Height_in", "Width_in", "Surface_Rect", "Diam_in", and "Surface_Rnd" because their information can be conveyed by "Surface". Similarly, we remove "material", "mat", "lands_sc", "lands_elem", "lands_figs" becasue their information can be conveyed by "mat_recode" and "landsALL", respectively.

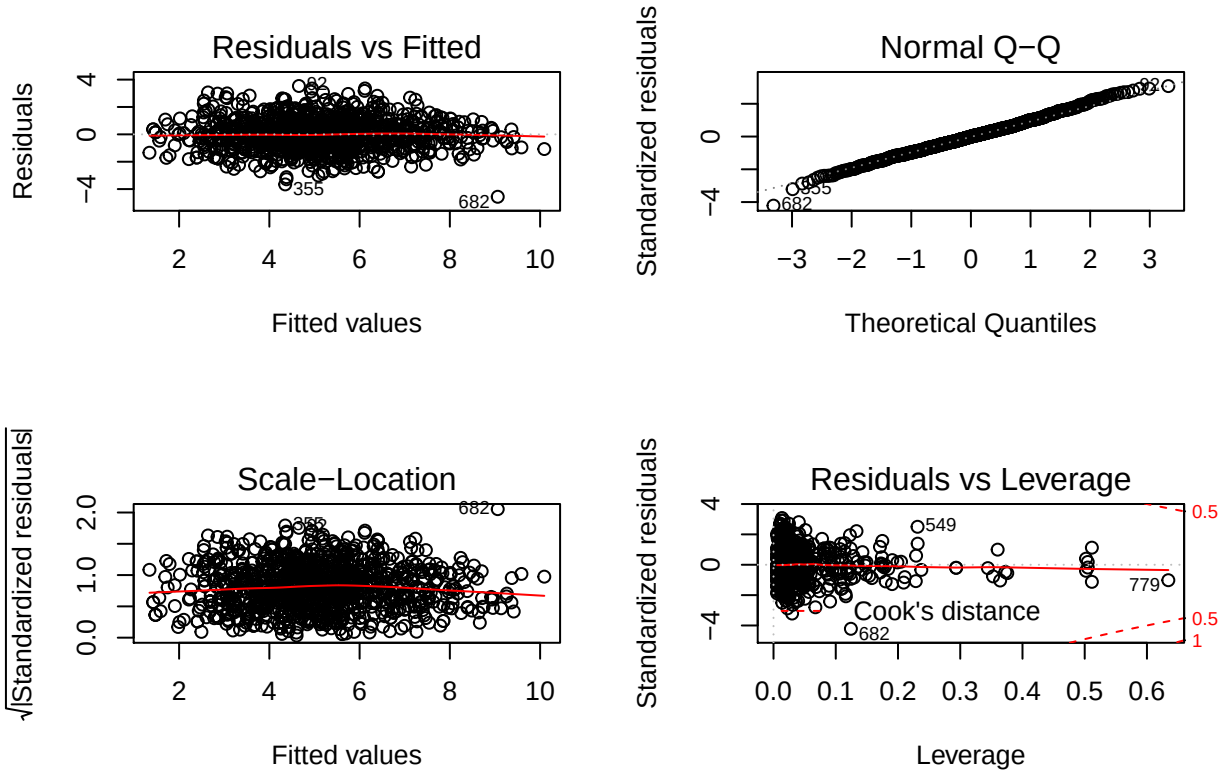
After all of the preparation, we run AIC and BIC analysis on our raw model, which includes all the variables remaining. In the end, we decide to use BIC analysis, and use the final model generated from BIC analysis as our first step to understand the dataset using linear models. The model obtained after BIC analysis is as follow:

"log(price) ~ dealer + year + school_pntg + diff_origin + artistliving + winningbiddertype + Surface + nfigures + engraved + prevcoll + othartist + paired + finished + lrgfont + othgenre + discauth"

After this, as we use the model obtained from BIC analysis to study the dataset, we slowly remove some of the variables, such as "othartist" and "nfigures", which we believe do not contribute much in our model. Later on, we also study the interaction effect between the predictors ("winningbiddertype:prevcoll"). We even go back to our initial model and manage to find an important interaction term between a predictor and a variable that seems insignificant on its own (Surface:mat_recode). Eventually, we finalize our model as follow:

"log(price) ~ dealer + year + school_pntg + diff_origin + artistliving + winningbiddertype + Surface + engraved + prevcoll + paired + finished + lrgfont + othgenre + discauth + winningbiddertype:prevcoll + Surface:mat_recode"

**Residual**




As we observe from the residual plots, we see a straight horizontal line, around which there is a patternless cloud of points. This shows that the independence assumption of linear model is well followed.

**Table of coefficients and CI**

| | Estimate | 2.5 % | 97.5 % |
|---|---|---|---|
| (Intercept) | -150.98 | -181.78 | -120.19 |
| dealerL | 1.57 | 1.28 | 1.87 |
| dealerP | 0.44 | 0.05 | 0.82 |
| dealerR | 1.78 | 1.53 | 2.03 |
| year | 0.09 | 0.07 | 0.10 |
| school_pntgF | -0.69 | -0.86 | -0.52 |
| school_pntgI | -0.70 | -0.93 | -0.48 |
| school_pntgOTHER | -0.88 | -1.92 | 0.15 |
| school_pntgX | -1.09 | -1.75 | -0.44 |
| diff_origin | -0.66 | -0.85 | -0.48 |
| artistliving | 0.33 | 0.08 | 0.57 |
| winningbiddertypeB | 1.14 | 0.44 | 1.84 |
| winningbiddertypeBC | 1.60 | 0.84 | 2.36 |
| winningbiddertypeC | 0.85 | 0.54 | 1.15 |
| winningbiddertypeD | 0.81 | 0.55 | 1.06 |
| winningbiddertypeDB | 1.79 | 0.14 | 3.45 |
| winningbiddertypeDC | 2.17 | 1.80 | 2.54 |
| winningbiddertypeDD | 1.79 | 0.82 | 2.77 |
| winningbiddertypeE | 0.37 | 0.04 | 0.69 |
| winningbiddertypeEB | 0.19 | -1.44 | 1.82 |
| winningbiddertypeOTHER | 2.26 | 0.60 | 3.91 |
| winningbiddertypeEC | 1.57 | 1.09 | 2.06 |
| winningbiddertypeU | 0.48 | 0.19 | 0.78 |
| Surface | 0.00 | 0.00 | 0.00 |
| engraved | 0.53 | 0.20 | 0.86 |
| prevcoll | 0.77 | 0.01 | 1.52 |
| paired | -0.30 | -0.45 | -0.14 |
| finished | 0.65 | 0.44 | 0.86 |
| lrgfont | 0.90 | 0.65 | 1.16 |
| othgenre | 0.47 | 0.23 | 0.71 |
| discauth | 0.50 | 0.20 | 0.80 |
| winningbiddertypeC:prevcoll | 0.11 | -1.02 | 1.23 |
| winningbiddertypeD:prevcoll | -0.04 | -0.92 | 0.85 |
| winningbiddertypeDC:prevcoll | -1.05 | -2.14 | 0.04 |
| winningbiddertypeE:prevcoll | 0.29 | -0.94 | 1.51 |
| winningbiddertypeEC:prevcoll | 0.06 | -1.53 | 1.64 |
| winningbiddertypeU:prevcoll | -1.90 | -4.30 | 0.51 |
| Surface:mat_recodemetal | 0.00 | 0.00 | 0.00 |
| Surface:mat_recodeother | 0.00 | 0.00 | 0.00 |
| Surface:mat_recodepaper | 0.00 | 0.00 | 0.00 |
| Surface:mat_recodeuncertain | 0.00 | 0.00 | 0.00 |
| Surface:mat_recodewood | 0.00 | 0.00 | 0.00 |

# 4. Conclusion and Summary

Based on the summary, in order to compute the (median) price for the "baseline" category, we set all factors to its reference level, while taking the median of individual continuous variables, and multiply them by their corresponding coefficients. Then, we take exponential to get from "log(price)" to "price".

```
## [1] 19.83399
```

From our summary, one of the important findings we would like to highlight is that "dealer" and "winningbiddertype" affect the price the most, as they have the largest estimate. This is expected, because different dealers may have the price differently, or even may have different sources of painting to sell, thus resulting in different price ranges (ie. some dealers may have sources of highly-sought paintings, whereas other dealers may just have normal paintings). The same reasoning applies to "winningbiddertype" as well. When a dealer on behalf of collector is buying a painting, he will definitely have a different ulterior agenda in mind from, say, a buy himself.

One of the potential limitations about our model is that we may not be able to consider every single possible combination of interation terms. In fact, as we analyze the dataset, we only consider interaction between two variables. However, interactions can occur even with multiple variables, and our model does not consider EVERY SINGLE of these potentially significant multiple interactions.

According to our summary, we believe that interactions are important. In our model, the most important variables are "finished", "prevcoll" and "lrgfont". To interpret how they influence the (median) price, we go back to previous section and obtain the confidence interval of coeffcients for both "dealer" and "winningbiddertype". Then, we take exponential of it so as to get from "log(price)" to "price".

(a). "finished"

```
##    2.5 %   97.5 %
## 1.558386 2.360873
```

The price of paintings noted for their highly polished finishing is 1.558386 to 2.360873 times of that of paintings not noted for their highly polished finishing.

(b). "precvoll"

```
##    2.5 %   97.5 %
## 1.014806 4.553613
```

The price of paintings whose previous owner is mentioned is 1.014806 to 4.553613 times of that of paintings whose previous owner is not mentioned.

(c). "lrgfont"

```
##    2.5 %   97.5 %
## 1.906738 3.198521
```

The price of paintings to which the dealer devotes an additional paragraph (in large font) is 1.906738 to 3.198521 times of that of paintings to which the dealer does not do so.