Multiple Linear Regression of Contemporary and Modern Art Auction Market in Hong Kong, 2016–2020

Summary: This project is to explore the predictive variables behind the auction prices of paintings. A special focus is on paintings of contemporary and modern art which were sold at major auction houses in Hong Kong from 2016 to 2020. We are interested in whether the location and time of auctions will affect the sale price. Also, whether the signature and medium of paintings will be important factors for modern art. It is discovered that the location is not a significant factor whereas the time matters. The signature and medium may also be important but more data are required to further examine these findings.

Introduction

One might think that the price of art may be unpredictable. However, it is actually necessary to provide a pre-sale estimate of all the works sold in the auction. Currently, humans are more accurate than machines in compiling these estimates. However, the process of manually evaluating artworks is slow, expensive, and limited by the number of human experts available. The regression model can potentially be extended to evaluate all artworks, thereby automating the pricing of all artworks on the market and off-site, thereby increasing buyer liquidity. To this end, the project aims to explore multiple linear regression models to price past auctions and automatically evaluate them when evaluating prices in the art auction market. This project is interested in whether the location and time of auctions will affect the sale price. Also, whether the signature and medium of paintings will be important factors for modern art. Other interesting facts are also explored in the discussion section.

Data

The data used for this analysis come from askART, an online artwork database which contains auction records dating back to 1987 and features over 350,000 artists[1]. This analysis is based on the auction records at Sotheby's and Christie's the two most well-known international auction houses in Hong Kong in the last 5 years. The original datasets contain artwork across the genre of contemporary and modern art, ranging from sculptures, prints, mixed media, to paintings. Only auction records of paintings are retained. The edition variable is therefore discarded since it is more related to screen prints and reproduction. For the same reason, the foundry variable is not kept as it is especially for sculptures.

Both Sotheby's and Christie's have established and adhered to the rules of so-called English auctions where the bidding starts low and increases as the bidders compete with higher bids [1]. Once the bidding ends, the auctioneer will hammer down the price, which is thereby called as the hammer price [1]. However, hammer price is usually not the final price buys finally pay for the artwork. Auction houses charge for a premium on top of the hammer price. The sale prices in this data include the hammer prices and premiums in dollars. Before the auctions, sellers will also set a reserve price, which determines the lowest price of the artwork and is completely concealed by the auction houses. The artwork remains unsold until the bidding reaches the reserve [1][2]. In this dataset, approximately 19% of the paintings are unsold. The auctioneer also provides a high estimate as well as a low estimate, which is at or above the

[1] www.askart.com

reserve price according to auction rules [1][2]. According to Ashenfelter [3], the average of low and high estimates is highly correlated with the final sales price. Therefore, both estimates in dollars are included for each auction lot in the data. A dummy variable is constructed to indicate whether the painting is sold or not.

EDA

With clean data, exploratory data analysis (EDA) was performed to detect patterns and potential associations among data. The distribution of our outcome variable, the sale price, was first examined. It is found that the distribution of auction prices is approximately normal after the logarithm distribution, as shown in Figure 1. This implies that multiple linear regression might be an appropriate model. In this case, the log-transformed auction price is regarded as the response variable in the regression model.

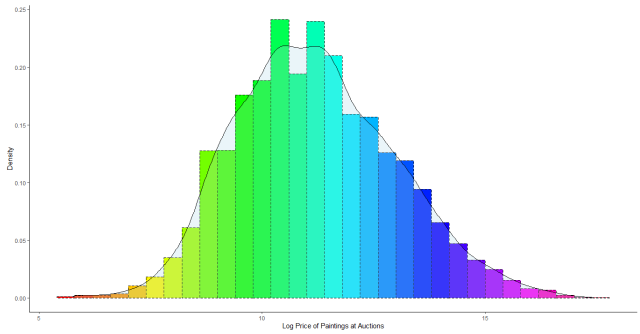

Figure 1: Distribution of log-transformed sale price

In this dataset, the covariates can be classified into numerical variables and categorical ones, among which the majority are dummy variables. There are only two numerical variables, which are the range of pre-sale estimation and the surface area of paintings. The range of pre-sale estimation shows an evident logarithmic relationship to the ultimate sale price. It is probably because the pre-sale estimation from appraisers of auction houses is highly accurate so that their range of estimation is closely related to the ultimate price. Therefore, if one of these two variables is on the logarithmic scale, the linear relationship between them will only be reflected when they are on the same scale, which is shown in Figure 2. Similarly, according to Figure 3, large paintings tend to be sold higher despite that over 90% of the paintings are less than 5 m^2, which constitutes a cluster.

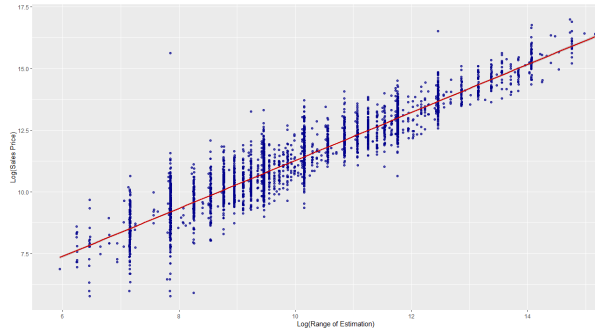

Figure 2: Scatter plot of log-transformed range of estimation versus log-transformed sale price with a fitted line.

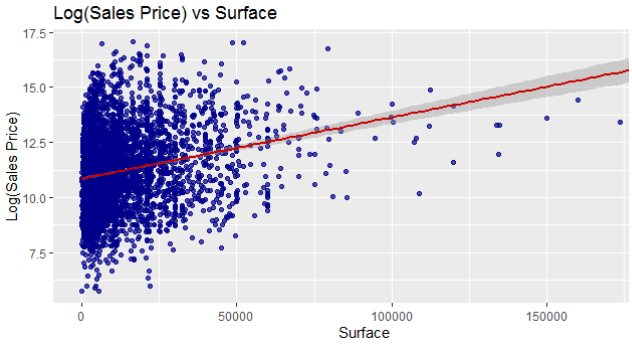

Figure 3: Scatter plot of surface area versus log-transformed sale price with a fitted line.

In terms of categorical variables, surprisingly, there is no difference in sales between the two auction houses in Hong Kong, Sotheby's and Christie's, during the last five years. In addition to the auction house, there are 20 categorical features in three groups, which are the artist's signature, the medium of paintings, and the auction time. Among these, variables related to the auction time appear to be significant. Auction houses tend to earn the most in September despite that there are no records in February and August. Also, Sotheby's has been on par with Christie's in the past four years. However, this year is an exception, the first time that Sotheby's is slightly behind Christie's. As for the medium, expensive materials such as oil seem to boost the price of artwork as well. Moreover, the advantages of signed paintings in the auction market are relatively weak compared to those unsigned. It is probably because the artworks in this dataset are all in the genre of modern art where an artist's signature is not that important compared to the style of old masters.

Model

Missing Data Imputation

It is worth noting that the clean dataset still contains missing values. To be specific, 4.25% of the paintings are not attached with the size information, which leads to missing values in the surface area variable. Likewise, 0.36% of the pre-sale estimates are not disclosed by auction houses. Accordingly, the range of estimation is incomplete. Even though the proportion of missing data is small, the number of observations is limited relative to a large number of variables, which is over 20. To avoid information loss, which potentially misleads to biased conclusions, multiple imputations were performed on this dataset with the assumption that data are missing at random (MAR).

Without loss of generality, 10 imputed datasets were generated with mice, which implements an iterative Markov Chain Monte Carlo algorithm. In particular, Bayesian linear regression was used to impute the range of estimation, which was restricted to positive values by post-processing. Whereas, classification and regression trees (CART) was applied for the surface area as a majority of categorical variables only contain two levels, which constitute binary trees naturally. With 2 outliers excluded, no trends exist in the mean and standard deviation of the imputed values at the later iterations. Therefore, this imputation method is acceptable for this dataset. Also, the distributions of imputed datasets closely approach the observed data, especially for the surface variable. The imputed range of estimation is shown on the log scale for clarity.

Model Selection

A complete dataset is randomly selected among imputed data and used for model selection. Firstly, a preliminary model was built on features that indicate strong relationship with the response variable in EDA. Stepwise regression was next conducted to select predictive variables iteratively until the regression converges at the final model. In this process, covariates were determined by ANOVA tests.

According to our exploratory data analysis, the two continuous variables, the range of estimation and the surface area, both showed positive linear relationship with the sale price and were therefore included in our preliminary model. Time-related predictors were also considered given our findings from EDA. Other than the aforementioned variables, there are a large number of dummy variables related to artists' signatures and medium of paintings. All the remaining variables, especially these dummy variables, were determined by stepwise regression with Akaike information criterion (AIC).

The stepwise regression is started with a null model which only considers the intercept as well as a full model that includes all the possible main effects and interactions. It is worth noting that not all of the

categorical variables can be examined for two-way interactions due to the lack of data. Variables in the model were further investigated by ANOVA tests at the 5% significance level. In this process, stepwise regression was refitted with the updated full model and finally converge at the following model, which was thereby considered our final model and formulated below. Our final model achieved an adjusted R-squared of 0.894 with most of the predictors significant at the 5% level. The more detailed summary of this model is attached in the appendix.

```
> summary(mbi261)

Call:
lm(formula = sales_price_log ~ estimate_range_log_c + surface_c +
    auction_weekday + auction_month + paper + lacquer + oil +
    wood + canvas + oil:titled + titled + auction_location, data = da)
```

Model Assessment

Linear regression assumptions of this model were evaluated to ensure the reliability of our inference and further prediction. With standardized residuals plotted against fitted values, it was noticed that 0.85% of the data were significantly out of the range of [-3, 3] and therefore regarded as outliers. We removed the outliers and refitted the model. As shown in Figure 4, all the assumptions are satisfied. It is because the residuals spread equally and independently around 0 in plots (a). Moreover, the normality assumption is largely satisfied as reported by the QQ-plot in (b). This model also meets the linearity assumption since there no pattern exists in the residual plots (c) and (d) versus each continuous variable.

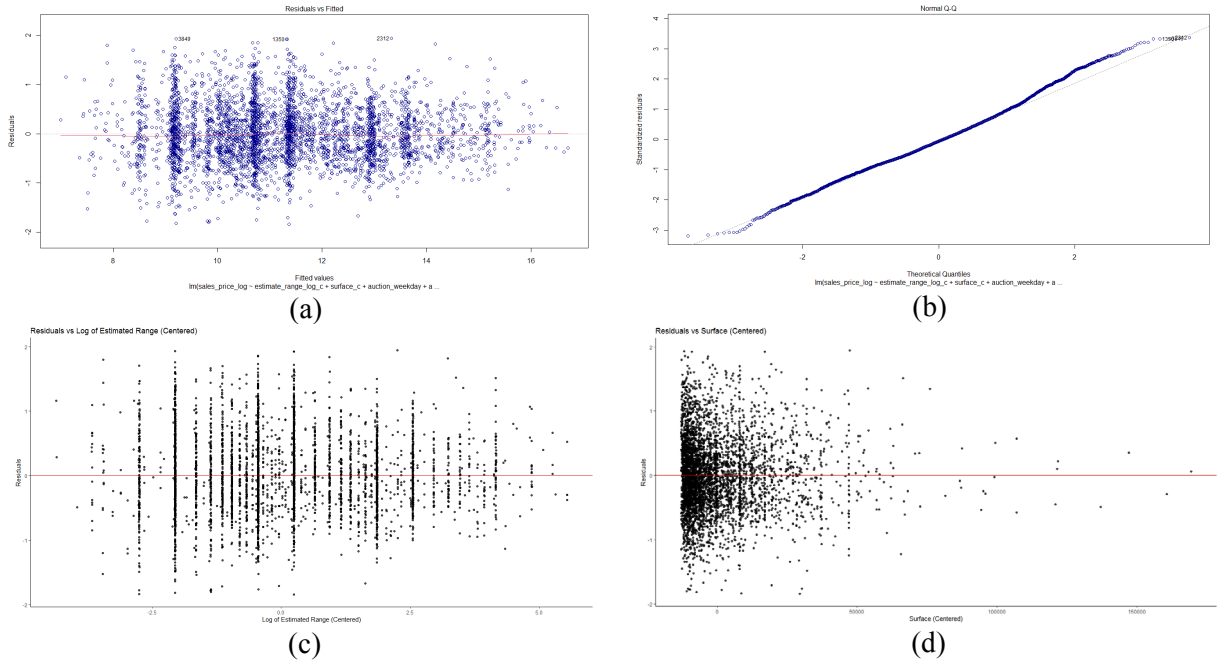

Figure 4: Assessments of linear regression assumptions: (a) Independence of errors and equal variance; (b) Normality; (c) Linearity of range of estimation; (d) Linearity of surface area.

Furthermore, no multicollinearity occurs in this model as the VIF scores of both continuous variables are satisfactorily below 5. In addition to excluded outliers, high leverage points exist but none of these observations are influential according to their Cook's distances.

Discussion

In this section, the results of our final model are interpreted to answer our initial inferential questions. Firstly, it is consistent with our EDA findings that whether a painting is sold at Christie's or Sotheby's makes no difference to the sale price in Hong Kong. However, the time of the auction in terms of weekdays and months are significant. Particularly, auctions on Saturday are expected to boost the sale price most in a week while auctions on Wednesday and Thursday are estimated to make the paintings less appealing. It might be counterintuitive that all months are expected to have negative effects on the sale price except for June. Similarly, most of the medium included in this model negatively affect the sale price except for lacquer. The reason may be that painting on lacquer is extremely difficult which requires not only artistic accomplishment but also crafting skills. Therefore, paintings on lacquer are rare and regarded to have high collection value, especially in Asian art. It is also because of the limited data that not all predictors can be sufficiently explained other than its statistical significance. Still, the signature variable is significant to the auction price. If a painting is titled by the artist, its price is expected to increase by $1.24.

In addition, the most significant predictors to the auction prices of a painting is the range of pre-sale estimates. This implies that paintings with larger range of estimates from the auction houses are expected to have higher sales price. The confidence interval of the coefficients at the log scale is attached in the appendix. Also, this model has a RMSE of 8.91 in k-fold validations, which is extremely close to the results from non-parametric methods such as random forests.

Conclusions

This model has explored the predictor variables to infer the prices of paintings at auction houses. To conclude, the range of estimations, time of auctions, medium, and signature are important factors. Most of the previous literature focuses on supervised or unsupervised learning methods to predict auction prices while little focus on specific predictors that drive the prices. It is worth noting that due to limited data a portion of interactions and main covariates are not able to explore in this study. In the future, it is aimed to explore further in this topic with web-scrapping techniques to obtain data and predict with more complex methodologies.

References

[1] O. Ashenfelter, "Art auctions," in *A Handbook of Cultural Economics*, no. November, 2003, pp. 32–39.

[2] A. Korteweg, R. Kräussl, and P. Verwijmeren, "Does it Pay to Invest in Art? A Selection-Corrected Returns Perspective," *Rev. Financ. Stud.*, vol. 29, no. 4, pp. 1007–1038, Apr. 2016, doi: 10.1093/rfs/hhv062.

[3] O. Ashenfelter, "How Auctions Work for Wine and Art," *J. Econ. Perspect.*, vol. 3, no. 3, pp. 23–36, 1989, doi: 10.1257/jep.3.3.23.

```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.138e+01  1.176e-01  96.785  < 2e-16 ***
estimate_range_log_c       9.554e-01  6.136e-03 155.710  < 2e-16 ***
surface_c                  6.423e-07  6.042e-07   1.063 0.287815
auction_weekdayMonday      2.843e-01  8.209e-02   3.464 0.000538 ***
auction_weekdaySaturday    3.034e-01  7.849e-02   3.865 0.000113 ***
auction_weekdaySunday      2.248e-01  7.842e-02   2.867 0.004170 **
auction_weekdayThursday   -1.047e-01  8.181e-02  -1.280 0.200673
auction_weekdayTuesday     2.768e-01  1.399e-01   1.978 0.047989 *
auction_weekdayWednesday  -1.376e-02  1.185e-01  -0.116 0.907562
auction_month3            -2.085e-01  1.001e-01  -2.083 0.037323 *
auction_month4            -3.998e-01  1.002e-01  -3.988 6.76e-05 ***
auction_month5            -5.009e-01  1.152e-01  -4.347 1.41e-05 ***
auction_month6             1.142e-01  1.271e-01   0.899 0.368907
auction_month7            -1.515e-01  1.257e-01  -1.205 0.228424
auction_month9            -2.161e-01  1.118e-01  -1.933 0.053249 .
auction_month10           -4.647e-01  1.004e-01  -4.627 3.82e-06 ***
auction_month11           -5.315e-01  1.162e-01  -4.574 4.91e-06 ***
auction_month12           -4.358e-01  1.651e-01  -2.639 0.008333 **
paper1                    -1.423e-01  4.110e-02  -3.462 0.000542 ***
lacquer1                   1.930e-01  6.337e-02   3.046 0.002332 **
oil1                      -1.707e-02  2.182e-02  -0.782 0.434127
wood1                     -1.209e-01  6.023e-02  -2.007 0.044817 *
canvas1                   -4.492e-02  2.576e-02  -1.744 0.081240 .
titled1                    2.182e-01  5.578e-02   3.912 9.30e-05 ***
auction_location1          9.075e-02  6.279e-02   1.445 0.148465
oil1:titled1              -1.936e-01  7.379e-02  -2.624 0.008719 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5778 on 4332 degrees of freedom
Multiple R-squared:  0.8946,    Adjusted R-squared:  0.894
F-statistic:  1470 on 25 and 4332 DF,  p-value: < 2.2e-16
```