Using Linear Models to Analyze Returns in Pop Art

Brewer Cappellin, A[1], Coleman, A[1]
Department of Statistics, University of California at Santa Cruz[1]

Abstract

For this project, we investigated the performance of Pop Art as an investment by exploring the meaningful characteristics that determine the price of an artwork. We performed a linear regression analysis using log of price as our response variable and ten predictors that uniquely describe each piece of art. We created a market price index to show the variations of Pop Art over the 2001-2012 period when considering the regression coefficients from the predictor "year" as a factor. Our price index results showed that during this period, the Pop Art market appreciated by 5.52% on average per year in real USD while our benchmark, the index for the S&P 500 (S&P), barely depreciated by 0.01% on average per year in real USD. These results showed that investing in Pop Art, during this time period, provides a higher return than investing in the stock market or low-risk corporate bonds.

KEY WORDS: pop art, linear regression, price analysis, linear models, investments, price index

1. Introduction

Over the last several decades, news stories containing accounts of fine art selling for staggering amounts of money continue to astound the public. We can see many of these notable instances in just the last decade of the Pop Art Movement. In January 2017, a work titled "Masterpiece" by Roy Lichtenstein sold for $165 million. In November 2015, another Lichtenstein piece titled "Nurse" sold for $95.4 million. In May 2015, an untitled artwork by Jean-Michel Basquiat sold for $110.5 million. In March 2010, a Jasper Jhones piece was sold for $110 million. Finally, in October 2008, an Andy Warhol piece titled "Eight Elvis" sold for $100 million. All of the aforementioned famous artists are included in our dataset along with many other artists whose art does not exceed the multimillion-dollar mark.

The growth of these multimillion dollar sales, the expansion of high-net-worth individuals, and the need for portfolio diversification in recent years have commanded a lot of attention to the art investment world. However, these high prices should not necessarily imply high returns for investments.

In order to assess the return on investment for the Pop Art market during this time period, we will perform a multiple regression analysis that will use price as our response and other variables as predictors. To select these other variables, we will execute in-depth exploratory data analysis (EDA). We will then build our model from these predictors. Finally, we will divide our complete data set into a training and test set to assess the predictive capability of our model.

As a comparison to our first model, we will remove the high-value artworks from our dataset and fit a second model to see if removing these large observations improves the predictive power of our model.

1.1 Prior Analysis

Two papers helped to influence our methods and analysis. The first paper, from Mei J, Moses M (1), did a similar analysis with artworks ranging from 1875 to 1999. They performed multiple linear regression on the data and calculated the returns. Their work concluded a real return of 4.9% per year and a higher rate of 8.2% per year after 1950. The second paper, Renneboog L, Spaenjers C (2), created an index for a database of over 1,000,000 artworks across multiple art movements sold from 1957 to 2007. Their result showS that art appreciated in value by 3.97% per year in real USD; a trend similar to that of corporate bonds.

With these two papers in mind, we performed an analysis only focused on the Pop Art movement while in a turbulent period for the stock market. Between the years 2000 and 2012, we had two important market crashes: the dot-com crash in 2000 and the Financial Crisis of 2007. This time period provided an interesting data set while also creating difficulties that we will explore in later sections.

2. Data

We are using a data set comprised of multiple art movements obtained from a web scrape conducted over the course of six months. Using the complete data set would be far outside the scope for this project. Therefore, we will be doing our analysis on a subset of this data set: the Pop Art Movement. With this subset, our data consists of 27,124 observations; having over 20 variables that contain characteristics of the art and sale information from 2000 to 2012.

To help reduce and understand the myriad of potential variables for our model, we referenced Renneboog L, Spaenjers C (2). When trying to predict the price of an artwork, we see that our variables can be classified under the categories shown in Table 1.

Table 1: Variables

Sale attributes	*Artwork attributes*	*Artist attributes*
Price	Material	Name
Auction house	Signed	Origin
Location sold	Dated	Vital status
Year sold	Area	
Month sold		

2.1 Data Manipulation

Before fitting a model, extensive data manipulation is needed. First, we removed all duplicate entries in our data. Using web scraping as a data collection method, we obtained 103 duplicate sale entries.

Next, we want to make sure that our data are aptly prepared to create a price index. For this index, we need to remove data from incomplete years. After removing the partial data from 2013, we are left with our final data set that contains 25572 observations.

Finally, we needed to check the proportion of repeated sales in our data set. Unfortunately, we found 3462 repeated sales; roughly 13% of out total observations. Because of this artifact, we no longer can assume that our errors are independent. We continue with the analysis while keeping in mind that this might affect fitting our model as well as our results.

Now that we have established a final dataset, we corrected the prices for inflation. Our prices must represent real values from their respective time period. We selected the latest date of the works sold and proceeded to correct all of the prices to real values.

2.2 Exploratory Data Analysis

After our data manipulation, we performed exploratory data analysis to help understand our variables while assessing the usefulness of any transformed variables. As a baseline, we initially fitted a model with the data, without any transformations, and found that the errors were not $NID(0, \sigma^2)$ with a constant variance. Therefore, we proceeded to apply a log transformation to both our two continuous variables: area and price.

With our newly transformed data, we fit a base model that contained all of the variables mentioned in section 2. This model was unable to estimate the coefficients of the different origin variables. We deduced that issue resulted from collinearity between artist names and artist origins; creating redundant information in our model. Therefore, we made sure to remove the variable that was the least statistically significant when selecting on our final model.

We then created different box plots of our factors against our log of price to assess the behavior of our data. Compared to our plots without a transformation on price, these plots (with the log of price) greatly improved; moving our predictor variables closer to our assumption criterion. This helped solidify our decision to keep the log of price as the transformation for our response variable.

Our next plot is the Log of Price vs the Log of Area shown in Figure 1. This plot depicts two different, unintentional clusters of artworks that may cause problems with the fit of our final model. In this plot, we see that the smaller cluster results from artworks that have an area of over 22,000 sq. inches. Compared to the bulk of our data, this small cluster is most likely a group of outliers and will create difficulties when making predictions with our final model.

The main group of points, centered around a price of $22,000, similarly begins to disperse as the price of artwork increases. This dispersion will also cause problems with our model, especially concerning the prediction accuracy of high-valued art. As we will see in our later analysis with our final model, most of these artworks valued over $350,000 will be underestimated.

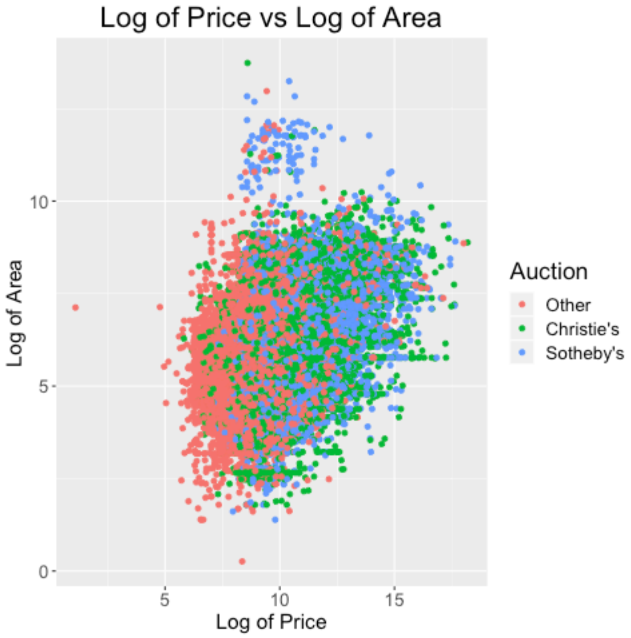

Figure 1: Scatter plot of log of Price vs log of Price

For the final portion of our exploratory data analysis, we created dendrograms of our factor variables in an attempt to diminish the number of levels in both origins of artist and auction location. These dendrograms helped us decide where to cluster our different levels.

In Figure 2 we can see how the clustering for the origin of the artist is dominated by Americans. We decided to group artist origin into two levels: “American-British” since the Pop Art Movement was predominantly an American-British movement, and “Other”.

Regarding Figure 3, we similarly clustered the place of sale, but into three levels for “New York”, “London”, which includes South Kensington in it’s larger metropolitan area, and “Other”.

Like these two dendrograms shown below, we clustered auction houses as well and reducing the levels to the large auction houses of "Sotheby's" and "Christie's", then including the rest in "Other".

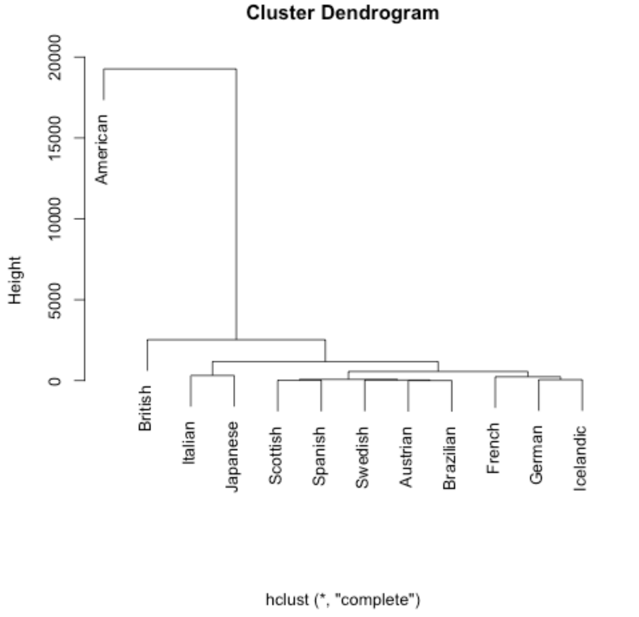

Figure 2: Dendogram for artist origin

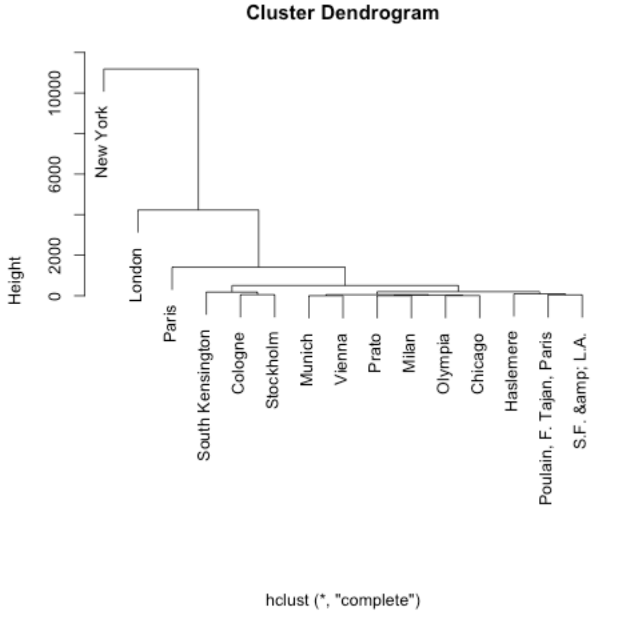

Figure 3: Dendogram for place of sale for the artwork

3. Methods

With our variables releveled, we can now begin to create a price index for the Pop Art Movement. We will use multiple linear regression to determine the value of the coefficients and the year variable to create the price index. We then use these index values to analyze returns per year. In addition to the construction of our price index, we will test the predictive power of our model. We will perform a 95:5 split of our data into a training and test set. Our training set will contain 24293 observations and the test set will contain 1279 observations.

3.1 Model

The equation of our model relates the log of the real price in USD to the year variable (factor), the log of the area (continuous), and the remaining variables that include artist traits, artworks characteristics, and sales information.

$$Ln(P_k) = \mu + \sum_{m=1}^{M} \beta_m X_{m,k} + \sum_{t=1}^{T} \gamma_t D_{t,k} + \alpha Ln(A_k) + \epsilon_k \quad (1)$$

We see P_k is the log of price of an artwork k, $X_{m,k}$ is the attribute m (Table 1) of artwork k, $D_{t,k}$ is the year variable for artwork k at year t and A_k is the variable for the log of area for each artwork k.

The coefficients β_m relate to the attribute m, coefficient α related to the log of the area, and γ_t are the coefficients for the year variable that will then be used to construct the price index for Pop Art.

The value of the index at year t will be calculated using

$$\Pi_t \equiv exp(\hat{\gamma}_t) \times 100 \quad (2)$$

Where Π_t is the value of the index at year t.

Because our initial year (2001) will be restricted as part of the intercept in our model, we must re-scale the year variable in order for it to be meaningful in the calculations of our index. This equation 2 provides us with the percentage growth for each year coefficient.

Silver M, Heravi S (3) provides us with an index that computes a geometric mean instead of an arithmetic mean. Because of the log transformation on the predictor variable, the geometric mean will be more useful in our case. Since the residuals for each year are normally distributed with constant variance, we corrected for the transformation bias using

$$\Pi_t^* \equiv exp(\hat{\gamma}_t + \frac{1}{2}(\hat{\sigma}_t^2 - \hat{\sigma}_0^2) \times 100 \quad (3)$$

where $\hat{\sigma}_0^2$ and $\hat{\sigma}_t^2$ are the estimated variances for the initial year (2001) and year t, respectively. Finally, we can calculate the estimated corrected returns for each year by

$$r_t^* \equiv \frac{\Pi_t^*}{\Pi_{t-1}^*} - 1 \quad (4)$$

3.2 Model Selection

With our model equation defined in equation 1 and our transformed variables, we fit three different models in order to determine the most meaningful variables for a final model. Using this final model, we will then create our price index and make predictions. We know from our exploratory data analysis that the artist name and artist origin cannot be included in the same model due to collinearity. Therefore, we need to determine which of the two variable is the most successful in predicting the log of the price.

For our first model, we included only artist name along with our other variables. We took note of the model selection criterion AIC, BIC, and R^2 to help decide between contending models.

Our second model removes the artist name and uses the origin variable with the rest of the variable. For this model, we also recorded our model selection criterion.

Finally, our third potential model uses the artist name while removing the origin and location of sale. We wanted to determine if the sale location contributed any predictive power to our model.

In Table 2 below, we summarized the model selection criteria for each of our three models. We see that by all three criterion, our first model performs the best. Therefore, we will keep our variables selected in section 2 and only remove artist origin in favor of artist name.

Table 2: AIC, BIC and R^2 for the three models fitted.

Model	*AIC*	*BIC*	R^2
Model 1	73979	74773	0.5508
Model 2	79720	80020	0.4282
Model 3	74336	75113	0.5441

3.3 Restrictions

We can see from our prior analysis and from Figure 4 that we have problems from heteroskedasticity (fat tails). Similarly, we know that not all of our necessary conditions are satisfied because of repeated sales: errors that are $NID(0, \sigma^2)$ with constant variance. Therefore, our model would need to be slightly changed or include different normalization of variables to achieve a better fit, which would be beyond the scope of this project. This restriction will thus affect our model fit and prediction accuracy.

4. Results

Table 3 shows the estimates for all the parameters in our final model. In this case, equation 1 is estimated using Ordinary Least Squares, where the dependent variable is the log of real prices in USD. Because the estimations have so many parameters, we will focus on their economic significance regarding their price prediction. The model sets restrictions to the following parameters by including them in the intercept: "Andy Warhol" for artist name, "other" for materials, "other" for auction house, "other" for locations, "dead" for vital status, "none" for signed, "none" for dated, "2001" for the year sold, and "May" for the month sold.

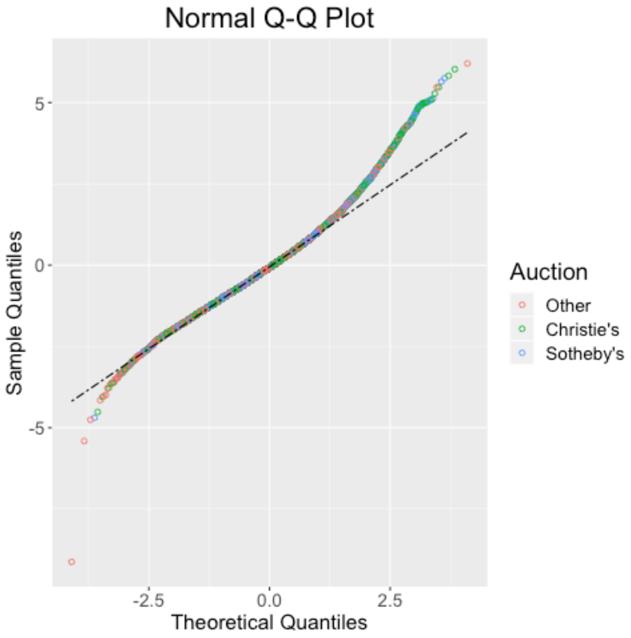

Figure 4: Normal Q-Q plot for Model 1, includes all original variables except origin of the artist

To estimate the price impact of the parameters, we used the following equation

$$Price\ Impact \equiv \exp(\hat{\beta}_m) - 1 \tag{5}$$

Where β_m is the estimated coefficient for the parameter we want to analyze (Table 3).

The price analysis shows how the parameters impact the value of the artwork. Since our model was fitted with Andy Warhol as the baseline for artists, when considering only the impact of other artists, all of them have a negative impact on price except three that have non significant coefficients.

When we look at the materials, we can see that of other all the materials have a positive impact on the price artwork except prints when compared to our baseline. In the case of "oil" and "acrylic", we see that they have the highest price impact with 404.56% percent and 340.65% respectively.

If we compare the impact of the auction house and location of where the art was sold, we find a positive impact of the parameters over our baseline. Christie's has an positive price impact of 74.21% and Sotheby's of 77.83%. In the case of the place sold New York's price impact is 41.47% and London's is 69.49%.

As expected (because we did not include interaction effects), if the artist is alive at the time of the sale, we see a negative impact on the price. If the artwork is dated, it has a positive impact. On the other hand, a curious result from our regression model is that if the artwork is signed, it has a negative price impact of −19.18%.

Table 3: Ordinary Least Squares Estimates for our final model. The dependent variable is the log of price and the predictors are as mentioned in the model selection.

```
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 7.76470    0.05813 133.567  < 2e-16 ***
ArtistAdami Valerio        -1.61723    0.06086 -26.571  < 2e-16 ***
ArtistArcangelo Allan      -2.12924    0.17464 -12.192  < 2e-16 ***
ArtistArman                -1.16472    0.04844 -24.045  < 2e-16 ***
ArtistArtschwager Richard  -1.36593    0.12984 -10.521  < 2e-16 ***
ArtistBarker Clive         -2.31578    0.24611  -9.410  < 2e-16 ***
ArtistBlake Peter          -1.01874    0.11922  -8.545  < 2e-16 ***
ArtistBoshier Derek        -3.66917    0.20166 -18.194  < 2e-16 ***
ArtistBritto Romero        -2.55985    0.23667 -10.816  < 2e-16 ***
ArtistCaulfield Patrick    -1.28412    0.12641 -10.159  < 2e-16 ***
ArtistChamberlain John     -1.08737    0.18968  -5.733 1.00e-08 ***
ArtistDine Jim             -1.16181    0.06530 -17.791  < 2e-16 ***
ArtistEggleston William     0.15338    0.07946   1.930 0.053575 .
ArtistErro                 -1.94397    0.06971 -27.887  < 2e-16 ***
ArtistFahlstrom Oyvind     -1.04432    0.15545  -6.718 1.88e-11 ***
ArtistGoode Joe            -2.91786    0.29069 -10.038  < 2e-16 ***
ArtistGrooms Red           -2.14149    0.15643 -13.689  < 2e-16 ***
ArtistHains Raymond        -0.42184    0.09279  -4.546 5.49e-06 ***
ArtistHamilton Richard     -0.53801    0.08178  -6.579 4.83e-11 ***
ArtistHaring Keith         -0.77173    0.03747 -20.597  < 2e-16 ***
ArtistHockney David        -0.62332    0.05421 -11.497  < 2e-16 ***
ArtistHopper Dennis        -0.97083    0.22321  -4.349 1.37e-05 ***
ArtistIndiana Robert       -1.08955    0.08181 -13.318  < 2e-16 ***
ArtistJohns Jasper         -0.20154    0.06067  -3.322 0.000896 ***
ArtistJohnson Ray          -0.69716    0.14248  -4.893 1.00e-06 ***
ArtistJones Allen          -1.91816    0.12217 -15.700  < 2e-16 ***
ArtistKatz Alex            -1.24363    0.07775 -15.995  < 2e-16 ***
ArtistKienholz Edward      -1.41704    0.18838  -7.522 5.57e-14 ***
ArtistKitaj R B            -1.22956    0.15075  -8.156 3.62e-16 ***
ArtistKlapheck Konrad      -0.40999    0.17664  -2.321 0.020294 *
ArtistKogelnik Kiki        -1.14230    0.21045  -5.428 5.75e-08 ***
ArtistKrushenick Nicholas  -2.63469    0.22290 -11.820  < 2e-16 ***
ArtistKusama Yayoi         -1.02406    0.07108 -14.407  < 2e-16 ***
ArtistLaing Gerald         -1.27372    0.23816  -5.348 8.96e-08 ***
ArtistLichtenstein Roy     -0.49553    0.02890 -17.146  < 2e-16 ***
ArtistLindner Richard      -0.87811    0.14082  -6.236 4.57e-10 ***
ArtistMax Peter            -2.73232    0.11786 -23.183  < 2e-16 ***
ArtistMurakami Takashi     -0.68087    0.08525  -7.987 1.44e-15 ***
ArtistNara Yoshitomo       -0.64444    0.07473  -8.623  < 2e-16 ***
ArtistOldenburg Claes      -0.90673    0.10179  -8.908  < 2e-16 ***
ArtistOpie Julian          -1.04903    0.10947  -9.583  < 2e-16 ***
ArtistPaolozzi Eduardo     -2.04875    0.13055 -15.694  < 2e-16 ***
ArtistPhillips Peter       -2.40149    0.18585 -12.921  < 2e-16 ***
ArtistPolke Sigmar          0.14585    0.05817   2.507 0.012180 *
ArtistPsaier Pietro        -3.00038    0.05944 -50.473  < 2e-16 ***
ArtistRamos Mel            -0.51958    0.12415  -4.185 2.86e-05 ***
ArtistRauschenberg Robert  -0.77017    0.05376 -14.325  < 2e-16 ***
ArtistRivers Larry         -1.43877    0.08544 -16.840  < 2e-16 ***
ArtistRizzi James          -2.87591    0.24720 -11.634  < 2e-16 ***
ArtistRosenquist James     -1.02036    0.08657 -11.787  < 2e-16 ***
ArtistRuscha Ed            -0.02035    0.06521  -0.312 0.754941
ArtistSaint Phalle Niki de -1.06665    0.08662 -12.314  < 2e-16 ***
ArtistSaul Peter           -1.39106    0.15596  -8.919  < 2e-16 ***
ArtistScharf Kenny         -1.80904    0.10116 -17.882  < 2e-16 ***
ArtistSegal George         -2.07362    0.18861 -10.994  < 2e-16 ***
ArtistSelf Colin           -2.37241    0.22654 -10.472  < 2e-16 ***
ArtistSmith Richard        -3.38927    0.17927 -18.906  < 2e-16 ***
ArtistTakano Aya           -1.40472    0.14693  -9.561  < 2e-16 ***
ArtistThiebaud Wayne        0.09950    0.07406   1.344 0.179110
ArtistTilson Joe           -2.76237    0.12559 -21.995  < 2e-16 ***
ArtistValdes Manolo        -0.53340    0.12790  -4.170 3.05e-05 ***
ArtistWesley John          -0.92177    0.13145  -7.013 2.40e-12 ***
ArtistWesselmann Tom       -0.51927    0.03505 -14.815  < 2e-16 ***
MaterialACRYL               1.48309    0.02838  52.267  < 2e-16 ***
MaterialOIL                 1.61851    0.03245  49.878  < 2e-16 ***
MaterialPRINTS             -0.70246    0.02247 -31.268  < 2e-16 ***
MaterialWRKSPPR             0.19835    0.02439   8.133 4.40e-16 ***
AuctionChristie's           0.55507    0.02364  23.481  < 2e-16 ***
AuctionSotheby's            0.57568    0.02388  24.110  < 2e-16 ***
PlaceLondon                 0.52763    0.02807  18.797  < 2e-16 ***
PlaceNew York               0.34694    0.02469  14.052  < 2e-16 ***
AreaLN                      0.31694    0.00591  53.627  < 2e-16 ***
AliveYES                   -0.15392    0.04320  -3.563 0.000367 ***
SignedYES                  -0.21289    0.02094 -10.164  < 2e-16 ***
DatedYES                    0.36217    0.01907  18.991  < 2e-16 ***
Year2002                    0.08481    0.04644   1.826 0.067833 .
Year2003                    0.15128    0.04557   3.320 0.000902 ***
Year2004                    0.30558    0.04398   6.949 3.78e-12 ***
Year2005                    0.36260    0.04193   8.647  < 2e-16 ***
Year2006                    0.56658    0.04117  13.763  < 2e-16 ***
Year2007                    0.84280    0.04070  20.710  < 2e-16 ***
Year2008                    0.54895    0.04354  12.608  < 2e-16 ***
Year2009                    0.45642    0.04623   9.874  < 2e-16 ***
Year2010                    0.59008    0.04297  13.732  < 2e-16 ***
Year2011                    0.61457    0.04308  14.266  < 2e-16 ***
Year2012                    0.47354    0.04208  11.253  < 2e-16 ***
MonthAPR                   -0.57735    0.02879 -20.053  < 2e-16 ***
MonthAUG                   -0.61866    0.14853  -4.165 3.12e-05 ***
MonthDEC                   -0.41564    0.03305 -12.576  < 2e-16 ***
MonthFEB                   -0.41855    0.03795 -11.029  < 2e-16 ***
MonthJAN                   -0.75660    0.07815  -9.681  < 2e-16 ***
MonthJUL                   -0.63590    0.04952 -12.843  < 2e-16 ***
MonthJUN                   -0.35275    0.03017 -11.693  < 2e-16 ***
MonthMAR                   -0.66855    0.03558 -18.790  < 2e-16 ***
MonthNOV                   -0.06112    0.02411  -2.535 0.011242 *
MonthOCT                   -0.59619    0.02563 -23.259  < 2e-16 ***
MonthSEP                   -0.81044    0.03493 -23.202  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.107 on 24196 degrees of freedom
Multiple R-squared:  0.5508,Adjusted R-squared:  0.549
```

4.1 Index

We can see the index we calculated for Pop Art alongside the index for the S&P from 2001 to 2012 in Figure 5. We used equation 3 to calculate the value of the Pop Art index and normalized both indices to start with an index value of 100 in 2001. For this comparison, we used the inflation-adjusted S&P, obtained from Multpl.com (4). However, for a complete, comprehensive analysis, we needed to similarly adjust for inflation of the art prices before comparing to the S&P index.

It is clear that during this time period, the Pop Art market out-performed conventional investments, even when comparing returns to low-risk corporate bonds. Both markets suffered from high volatility due to the aforementioned crashes, but the Pop Art market was more robust than the traditional stock market. An interesting result from our analysis showed that the dot-com crash had a small impact compared to the later Financial Crisis.

Another noteworthy trend, displayed clearly in Figure 5, shows a delay in the Pop Art market by about one year when compared directly to the stock market. This shows that people who invest in the Pop Art market have a slightly slower reaction time to alter their investments by about one year. This phenomenon can be explained by the time of the year in which large auctions occur. For example, in 2007, most of the sales closed before the bubble of the Financial Crisis burst, thus leaving the Pop Art market unaffected for that year, as seen below.

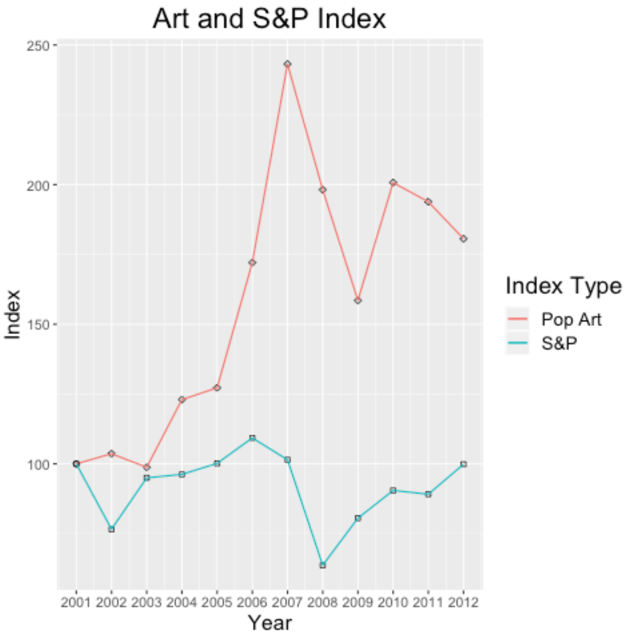

Figure 5: Plot over time for Corrected index and S&P

Table 4 allows us to analyze these results in much greater depth. During the aftermath of the dot-com crash, the Pop Art market remained mostly unaffected. However, the stock market suffered greater losses during the year 2002. Following this abysmal crash, both markets swiftly recovered; the stock market first by 2003 and then the Pop Art market by 2004. Later, leading up to the Financial Crisis, both markets suffered from high speculation, meaning that assets were continually purchased at higher and higher prices with the hope that their value will also increase in the near future.

We can see that the Pop Art market experienced more severe side effects from the crash by looking at the increase from 35.2% to 41.3% in 2006 and 2007 respectively. In the stock market, we only see an increase of 9.11% in 2006 and then a gradual to a sharp decline by 2007. This led to a big crash for both markets following the bubble burst in 2008. We can see that the Pop Art market declined 18.5% while the stock market decreased a staggering 37.2%. This disconnect continues when in 2009, the Pop Art market remained in decline at a rate of 20.0% while the stock market began its recovery at a rate of 26.5%. The Pop Art market did not begin to recover until the following year when returns in 2010 reached 26.6%. For the final years of our time period, 2011 to 2012, we see that the stock market continued its recovery while the Pop Art market remained in decline.

In conclusion, both markets suffered from high volatility during the period analyzed. However, despite this tumultuous time period, the Pop Art market clearly outperformed the stock market. Table 4 summarizes the returns for both Pop Art and S&P and also includes the Time Weighted Rate of Return (TWRR) over the 12 years period. We see that the Pop Art investments had an average return of 5.52% per year in real USD and the stock market returns were neutral at an average return of −0.01% per year in real USD.

Table 4: Returns for the S&P and Pop Art by year. The last row shows the Time Weighted Rate of Return for each index.

Year	Pop Art Returns (%)	S&P Returns (%)
2001	NA	NA
2002	3.63	-23.4
2003	-4.70	24.0
2004	24.5	1.31
2005	3.43	4.09
2006	35.2	9.11
2007	41.3	-7.16
2008	-18.5	-37.2
2009	-20.0	26.5
2010	26.6	12.3
2011	-3.42	-1.48
2012	-6.83	12.0
TWRR	5.52	-0.01

4.2 Prediction

We will now use the remaining 5% of our data, preserved in a test set from our initial split, in order to assess our model's prediction capabilities. One method that we used to determine our model's predictive power was to compare the correlation between the actual values of the out of sample (test) data and the predicted values. We obtained a value of 0.75, which indicates fair predictability for our model.

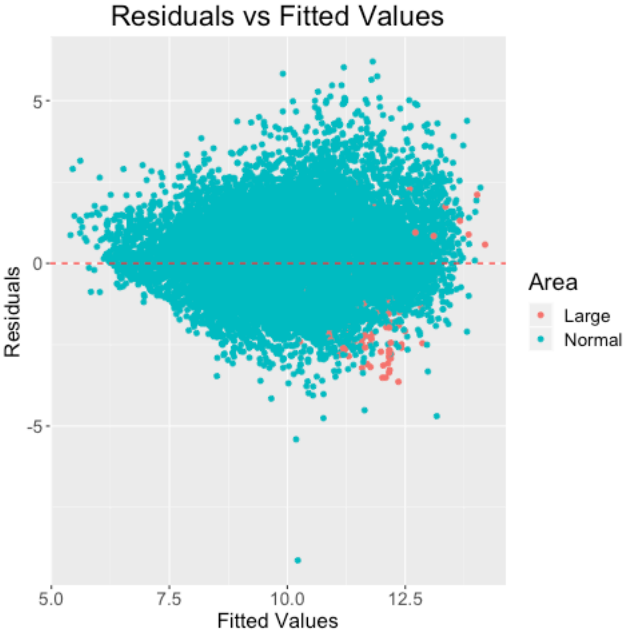

Figure 6: Residual plot for Model 1 colored by Area Size, where large is an area over 22,000 sq. inches.

As another method of predicting accuracy is to compare the Mean Squared Errors (MSE). For the final model, we obtained 1.2255 and for our out of sample MSE, we got 1.2657. This small difference between both numbers (3.28%) tells us that the out of sample prediction is performing similar to our final model.

As mentioned during EDA, the two different clusters that we observed in Figure 1 might give problems for our predictions. Therefore Figure 6 shows that the large artworks deviate in their residuals from the rest. It can be seen that the majority of the residuals are negative and far away from zero. This indicates that predictions for artworks with larger areas will not be as accurate.

Figure 7 shows the residuals for the predicted values from the out of sample test data set. This plot has a clear pattern that confirms what was explored in the model restrictions. We can see that our model has issues predicting high-valued artworks: as the price of the art goes up, the residuals increase, telling us that the model is under predicting the value of these artworks. Similarly, we can also see that our model had difficulty predicting for large area artworks, as hypothesized earlier. Here, most of our residuals fall below zero, indicating that the prices are mostly overpredicted.

Given these egregious errors in prediction, we will introduce an additional model in the next section where we removed high valued artworks and compare the difference in indices, returns, and predictions.

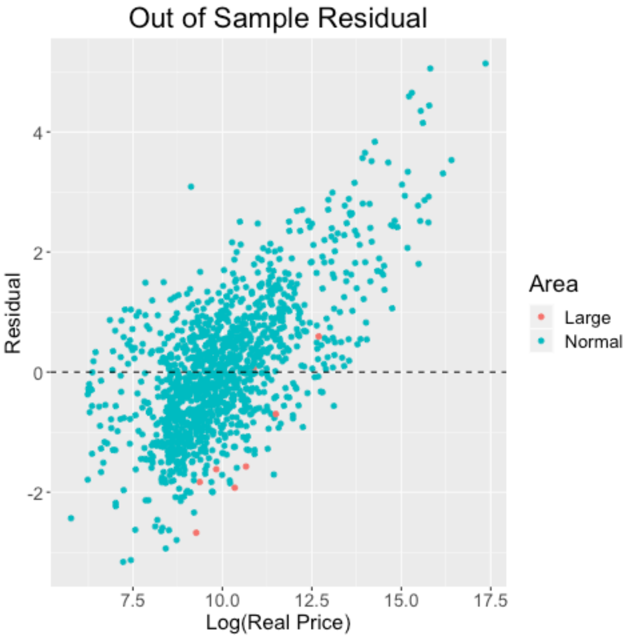

Figure 7: Plot of out of sample residual

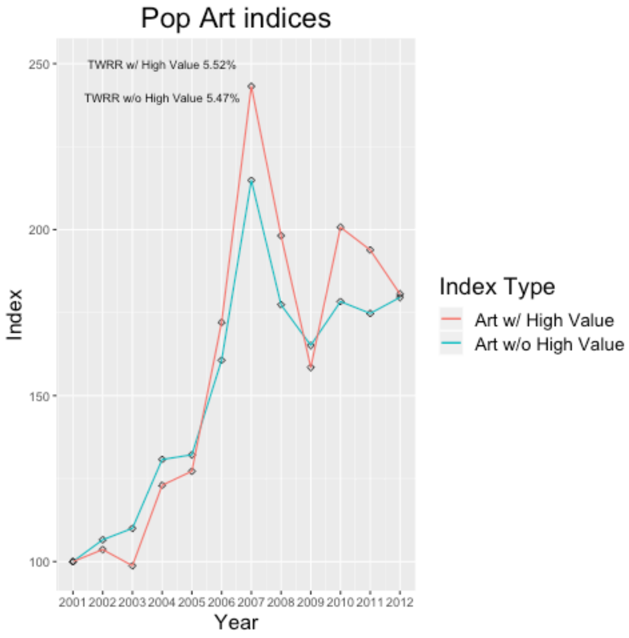

Figure 8: Plot over time for the Pop Art index with and with out High Value artworks

4.3 Additional Model

Since our final model possesses issues predicting high-value artworks, we created an additional model using the same variables but with excluding the high-value artworks from our data. We will only include artworks below $300,000. In total, 2012 observations were removed, or approximately 7.87% of the total data set.

Figure 8 compares both of the indices from the original model, with our whole data set, and this new model, using a subset of the data. As expected, both indices follow a very similar pattern, despite intersecting at various points in time. Most importantly, we observed no significant difference in the rate of returns. The original index had a 5.52% average yearly return in USD and the reduced data set had 5.47% average yearly return in USD.

Given that the indices are almost identical, we must now decide whether the predictions made by our addition model exceed those made by our final model. We once again computed and compared the Mean Squared Error. We found that the in sample MSE for the additional model is 0.8474 and the out of sample MSE is 0.8992. The small difference between both numbers (6.12%) tells us that the out of sample prediction is also performing well.

Figure 9 shows the out of sample residuals for our additional model. Unfortunately, we still are observing the same problem with accurately predicting prices for large artworks. However, we do see that are residuals and MSE are smaller as a direct result of removing high valued artworks.

Even though this additional model is incomplete in terms of developing a model for the entirety of Pop Art, we do see an improvement in the predictive accuracy of our model. Similarly, we see only a small difference in the market index and average yearly returns. Despite this improvement, our additional model still exhibits the same problems as our final model when predicting pricey artworks.

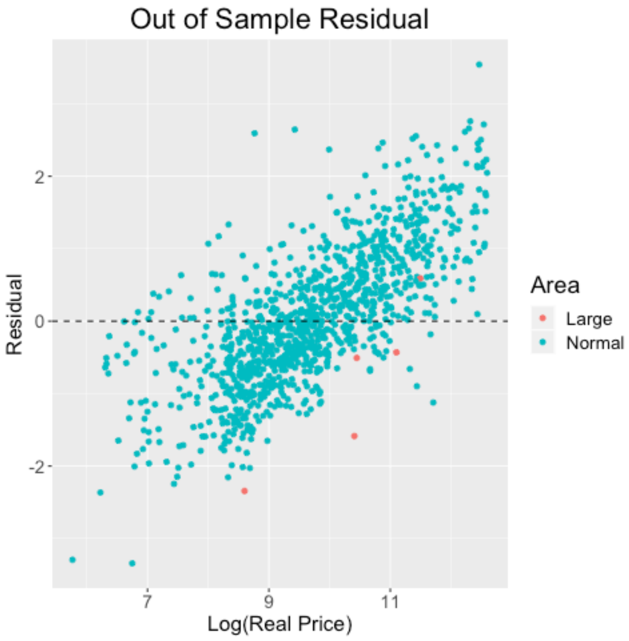

Figure 9: Plot of out of sample residual for data set with out High Value art

To summarize, the model only works better if we restrict the price range to below $300,000. This makes for an inefficient model since we retain the same issues while restricting the usefulness of our price index.

5. Conclusions

Given the growing popularity of investors to diversify their portfolios through non-conventional assets, we set out to analyze the return of Pop Art and compare it to traditional stock market trends. To execute this ambitious question, we investigated the Pop Art market by applying a multiple linear regression to a data set of over twenty-five thousand observations for over a decade of sales. Our price index indicates that Pop Art prices have increased on average by 5.52% per year in real USD between 2001 and 2012. This return is higher than results reported in prior works as well as in other movements of art. Most importantly, the results show a significantly greater return than the stock market during the same period.

With our final model for the complete training set, we tested its prediction capability using a test set. This analysis revealed problems with predicting high-value artworks as well as artworks with a large area. We observed that our residuals became larger as the price of the artwork went up. Most likely, a large portion of this phenomena could be explained by repeated sales of artworks. The existence of repeated sales causes our observations to be dependent, thus failing to meet our independent error assumption. Another likely cause could be from not having a roughly equal number of observations across our large price range. Without this extra data, all of our large values over $300,000 behave like outliers.

Finally, we fitted an additional model without high-value artworks to compare to our final model. Our additional model had lower in sample and out of sample MSE as well as smaller residual values when testing for predictions. However, we saw no significant significant difference in the average returns per year between both indices. The plot for the index of our additional model was very similar to the one from our final model. This led us to conclude that, despite the known problems with our original model, our index still painted a reasonably accurate picture of the Pop Art market. In conclusion, given our goal of analyzing price returns in the Pop Art market as a whole, our final model, using our complete training set, more appropriately satisfies our initial research question.

5.1 Future Work

Future works for this project can progress in many different directions. Collecting data from recent years could be used to compare to other investments in the current economic climate. The years available to us for this project coincide with two bubbles in the economy that makes establishing comparisons over an unsteady market unreliable.

Another direction would be to extend the index into other movements in art. This expansion in the scope would help to create a more inclusive and accurate representation of the art movement as a whole rather than the small subset of Pop Art.

With respect to the model, different variable transformations and interactions should be explored. For example, there may be an interaction between whether a piece is signed and if the artist is deceased. As portrayed in Figure 4, our necessary assumption of normality is not met since the Q-Q plot shows a large amount of data concentrated near the tails. Therefore, other transformations of the predicted variable may similarly solve this issue when including high-value art.

As a final consideration, future work should explore the possibility of including more variables into the analysis. Even though we included all of the important characteristics pertaining to each work of art, we did not eliminate the possibility that a combination of less important variables would have a stronger prediction accuracy. Either with adding or removing more variables, these explorations may make for a more robust model.

REFERENCES

(1) Mei J, Moses M (2002), "Art as an investment and the underperformance of masterpieces," *American Economic Review*, **92**, 1656–1668.

(2) Renneboog L, Spaenjers C (2012), "Buying Beauty: On Prices and Returns in the Art Market," *Management Science*, **ISSN 156-5501**.

(3) Silver M, Heravi S (2007), "Why elementary price index number formulas differ: Evidence on price dispersion," *J. Econometrics*, **140**, 874–883.

(4) Inflation Adjusted S&P 500 by Year (2018). Multpl.com. Retrieved 12 December 2018, *http://www.multpl.com/inflation-adjusted-s-p-500/table/by-year*