Hierarchical Multi-Level Modeling of Sold Paintings from Influential Auction Houses- A Linear Mixed Effects Approach

Co Sou

STAT 790: Case Seminar

Prof. Dana Sylvan

INTRODUCTION

This paper studies the modeling of characteristics that are believed to have an impact on the sold prices of paintings from some of the world's most well-known auction houses. To that end, a linear mixed effect model is used to quantify the relationship between the explanatory variables and the actual price realized by each painting. The ultimate goal is to gain some insights into how paintings are appraised and what characteristics contribute to their values. Researchers from various academic fields, ranging from Art History (Findlay, 2014) to Economics (Chanel, 1996), have attempted quantitative analysis of this topic. Here, we take a statistical framework in order to let us ask the question in a more fundamental manner.

First, we provide a description of the dataset and the motivation for using a mixed effect model. The data consists of 242,339 records of paintings sold from 22 of the world's most influential auction houses. These records were retrieved from two electronic databases commonly used by professionals in the industry, including auction house staffs as well as art historians and museum researchers (McNulty, 2013). Each record consists of the artist's name, title of the painting, lower and upper estimates, length and width, a Boolean variable for whether a painting is signed by the artist or not, the sales price realized, and the auction house where it was sold. Due to the nature of the auction process the model does not make use of the upper estimate since it is believed that the lower estimate has a significant influence on the sales price.

There are other intricacies about the data that need to be addressed. Firstly, only paintings are examined in this project, even though auction houses offer many products. Paintings are defined, in this study at least, to be two-dimensional objects that are made with a pigment and a physical support that lays the foundation. Therefore, certain items sold like portrait miniatures, certain contemporary artworks, and decorative art pieces are excluded from the dataset. Secondly, only records with full explanatory variables present are included. Missing data comes in multiple forms and it is argued that these records with missing values are Missing Not at Random (MNAR) (Tango, 2017) (Verbeke, 1997). For example, high profile paintings covered in the popular media are usually not given lower and upper estimates since that would conflict with the auction process. Traditional techniques like multiple imputation should not be employed on this type of missing data and hence, their records are completely omitted from study. That being said, the dataset consists of more than ample number of observations. Lastly, we make a note about (pseudo) replicates. A single painting might be bought from one auction house and then sold for a profit by an owner, which is then acquired by another auction house; this causes the number of observations to be superficially inflated. For this study only one unique painting is considered and whenever there are repeated measures, only the highest sales price record is included.

Finally, we justify the need for a linear mixed effect model. The question proposed here is: How does the choice of auction house influence the realized sales price of the paintings? We are interested in sampling an auction house from a pool of many, treated as a random variable, and see if there are effects on the prices based on the other explanatory variables treated as fixed effects. The question suggests a hierarchical multilevel model, in this case two levels, where the paintings and their fixed effects are nested within the auction house level variable (random effect).

Methodologies

We review some necessary model specifications and cite some works in the literature. We start with the form of the General Linear Mixed Effects Model for two-levels:

$$\boldsymbol{Y}_i = \boldsymbol{X_i\beta} + \boldsymbol{Z_i b_i} + \boldsymbol{\varepsilon}_i$$

where $\boldsymbol{b}_i \sim \boldsymbol{N}(\mathbf{0}, \mathbf{D})$, $\boldsymbol{\varepsilon_i} \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{\Sigma_i})$, and $\boldsymbol{b}_1, \dots, \boldsymbol{b}_n, \boldsymbol{\varepsilon_1}, \dots, \boldsymbol{\varepsilon_n}$ are all independent. Here, $\boldsymbol{Y}_i$ represents the vector of responses, $\boldsymbol{X_i}$ and $\boldsymbol{Z_i}$ matrices of covariates, $\boldsymbol{\beta}$ is a vector of the fixed effects, $\boldsymbol{b_i}$ is a vector representing the random effects, and $\boldsymbol{\varepsilon}_i$ is a vector of residual terms. The quantities $\mathbf{D}$ and $\boldsymbol{\Sigma_i}$ are covariance matrices; they are also called variance components (Searle, 2006).

The model specification follows:

$$\begin{aligned}\boldsymbol{SalesPrice_{ij}} &= \boldsymbol{\beta_0} + \boldsymbol{\beta_1\, Signed_{ij}} + \boldsymbol{\beta_2 Length_{ij}} + \boldsymbol{\beta_3\, Width_{ij}} \\ &+ \boldsymbol{\beta_4 LowEstimate_{ij}} + \boldsymbol{b_{0j}} + \boldsymbol{\varepsilon_{ij}}\end{aligned}$$

where $i \in \{1, \dots, 22\}$ indexes the auction houses, $j \in \{1, \dots, 242339\}$ indexes the paintings, $\boldsymbol{b_{0j}}$ represents the random effect intercept of the auction houses, which is assumed to be normally distributed with mean 0 and constant variance.

In the context of variance components, we use the Restricted Maximum Likelihood to get unbiased variance parameters because the MLE is biased (Corbeil, 1976). That being said, both methods can be used in linear mixed models when there is a large sample size, which is the case here (Jiang, 2017).

Another useful concept is the intraclass correlation coefficient, which describes how strong observations are correlated within their respective groups. This can be visually displayed to identify strong affiliations among auction houses and painting prices. This is given as:

$$\hat{\hat{\boldsymbol{\rho}}} = \frac{\hat{\boldsymbol{\tau}}^2}{\hat{\boldsymbol{\tau}}^2 + \hat{\boldsymbol{\sigma}}^2}$$

where $\boldsymbol{\tau^2}$ and $\boldsymbol{\sigma^2}$ are, respectively, the population variance between clusters and within clusters.

Finally, useful for the statistical inference of the random effects is the Empirical Bayes approach. This is a compromise between the pure Bayesian way of estimating the posterior distribution of the marginal distribution given as a prior and that of the Frequentist approach where we substitute the unknown parameters of the posterior with estimated values using the MLE or REML. This is used in predicting the random effects themselves and is given the name Empirical Best Linear Unbiased Predictors (EBLUPs), analogous to the term Best Linear Unbiased Estimators (BLUEs) in OLS from the Gauss-Markov Theorem. To be more specific, the values of the prior parameters are estimated by the Likelihood methods from the data themselves. For more information on this topic, see (Carlin and Louis, 2000).

Results

The data consists of 242,339 paintings, grouped according to their 22 auction houses from around the world. In this project, only painting sold under the price of $1.5 million USD are examined due to exploratory analysis showing that anything above to be ultra-high end. Figure 1A shows the distribution of the prices for each auction house, identified by informative abbreviations.

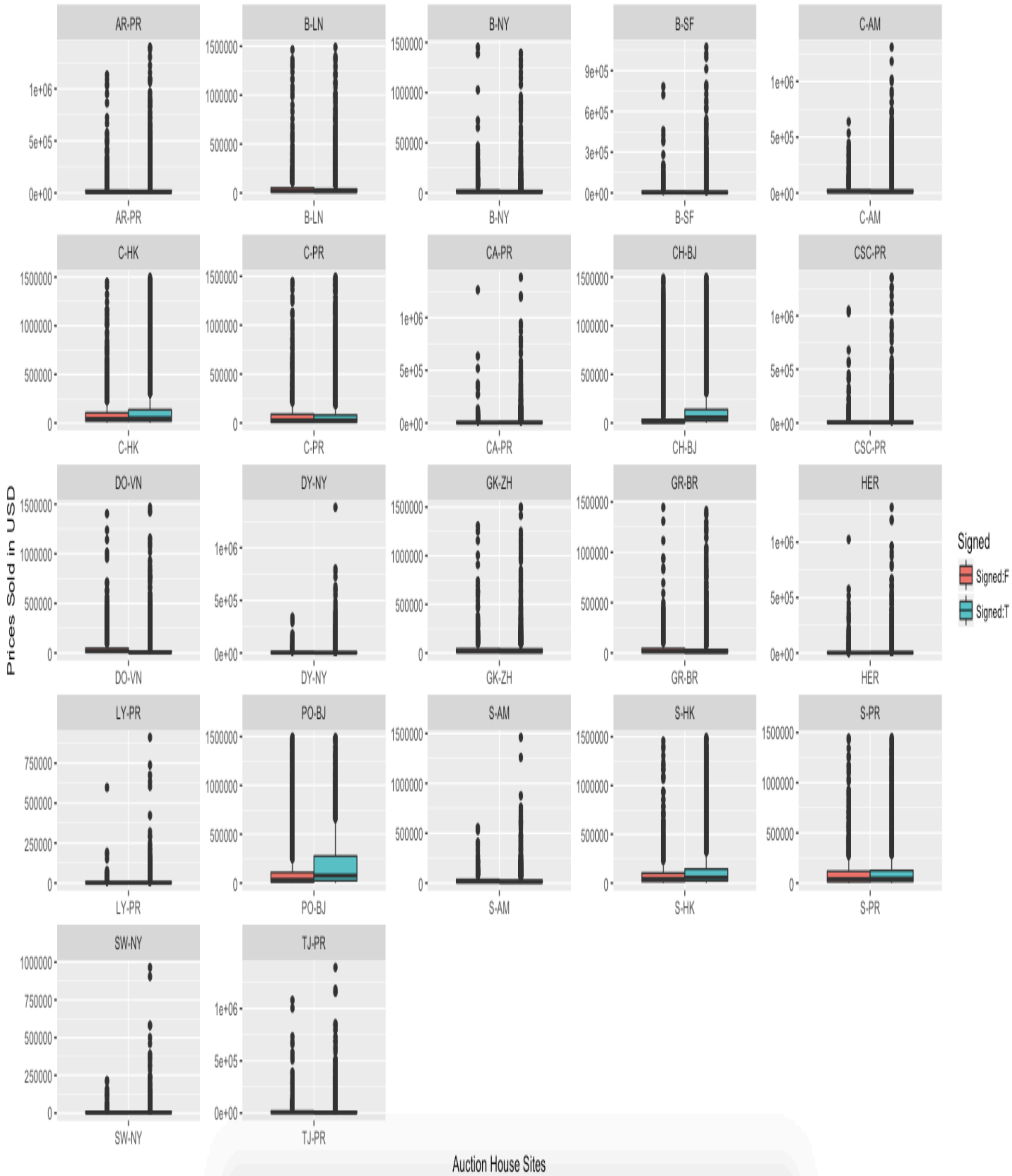

Figure 1A: Boxplots of prices realized for each auction house

We start with a brief discussion about the model selection process. It begins with fitting a model with only the main effects, followed by models with progressive complexity where more interactions are added until a saturated model is formed. Then we perform an ANOVA to compare these models, guided by p-values and AIC to choose the best one. Note that in order to perform the ANOVA we have to fit the models using the MLE so that the likelihood ratio test can be formed, instead of the REML. The use of the *nlme* R package supports this distinction (Pinheiro, 2000).

The following shows the diagnostic plots to ensure our model's assumptions are fulfilled. Below are the figures showing the residual plot and the QQ-plot for the fitted model. A log transformation has been applied to improve the normality. We see that the QQ plot in Figure 2A is reasonably normal. There exhibits some skewness at both ends of the tails but overall, most of the points lie on and around the linear line. In (Verbeke, 2009) it is argued that an imperfect normality plot might not suggest that the model is faulty and that we must look at the other diagnostic plots as well. Further examination of the extreme points shows that these are pieces with Chinese artists with large lengths and widths, most of which are calligraphy scrolls that have relatively medium sales price.

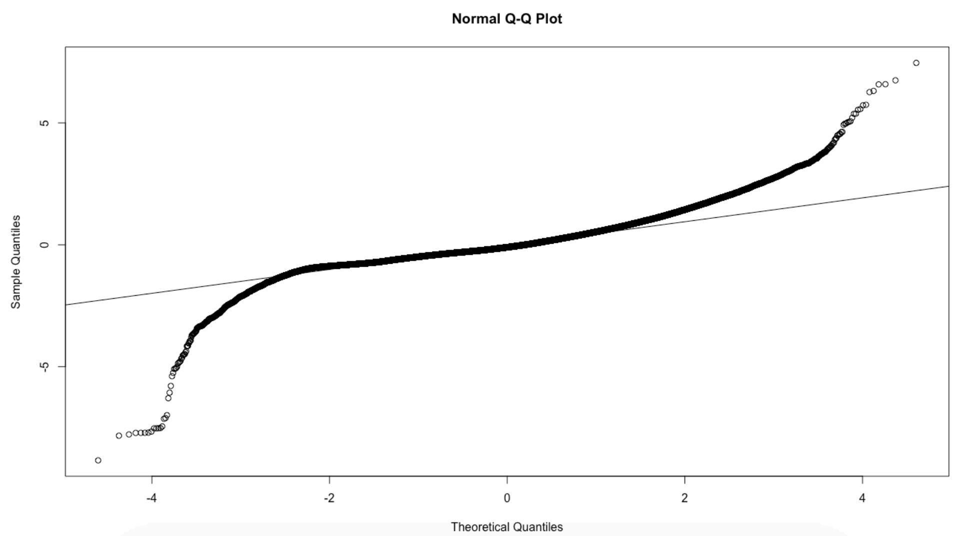

Figure 2A- QQ plot for normality for Mixed Effects Model

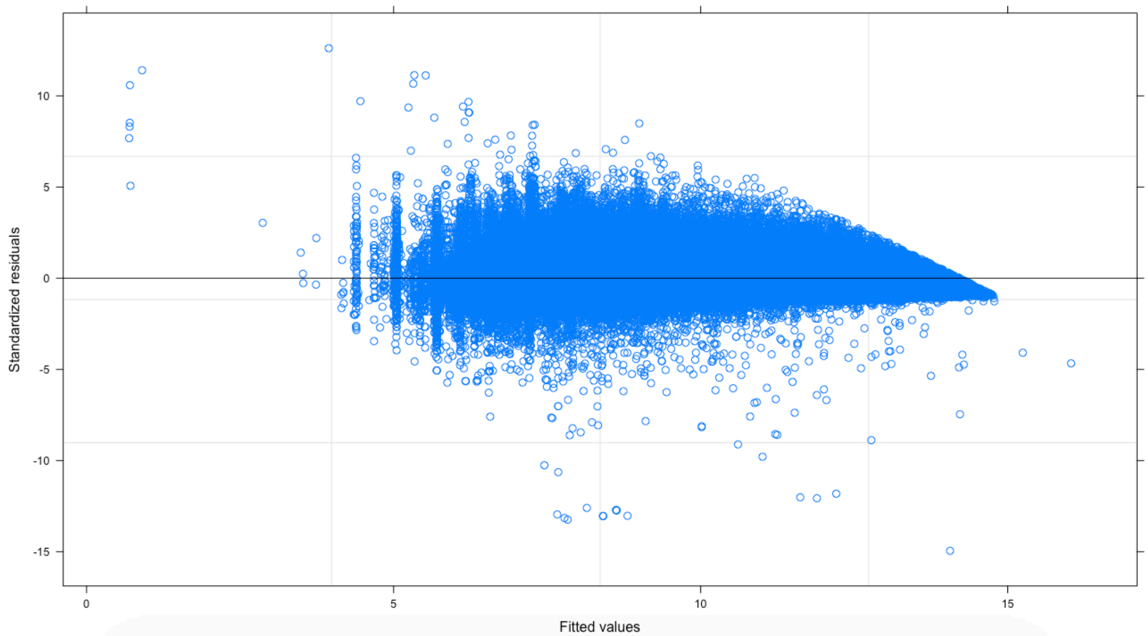

Figure 2B: Residual plot for Mixed Effects Model

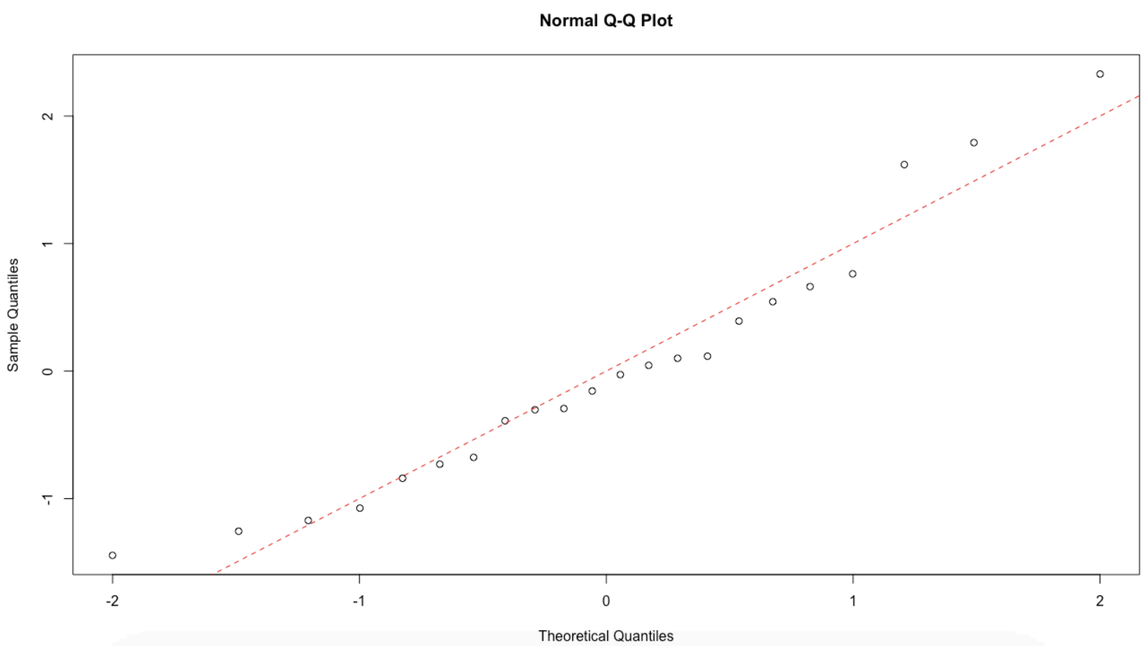

Figure 2C: QQ Plot for Normality of the Random Effects Model

Other diagnostics include the residual vs. fitted values plot in Figure 2B and the normality of the random effects in Figure 2C. We see that the residuals concentrate around the zero line and have a symmetric oval shape, which is a desirable trait. Figure 2C shows that the normality of random effects is also reasonably justified. There seems to be one auction house that deviates from this though; it turns out to be the San Francisco branch of Bonham's.

The regression output gives:

```
Linear mixed-effects model fit by REML
  Data: AuctionDataFiltered2
  Log-restricted-likelihood: -216784.9
  Fixed: form5
      (Intercept) log(LEstimateUSD)            Length             Width    SignedSigned:T
     0.9494408590      0.9446858228      0.0008438491     -0.0004549902     -0.0012238122

Random effects:
 Formula: ~1 | SiteID
        (Intercept)  Residual
StdDev:   0.2057266 0.5916618

Number of Observations: 242339
Number of Groups: 22
```

This gives the fixed effects model of:

$$SalesPrice = 0.94944 - 0.0012SignedTrue + 0.0008438Length - 0.0004545Width + 0.9446856LowerEstimate$$

This suggests that most of the Sales Price can be attributed to the lower price estimate given by the auction houses. It's confounding that a painting having a signature by the artists would slightly penalize the sales price even though intuition tells us that paintings with signatures symbolize authenticity, which would add to a painting's worth. Further examination of the fixed effects gives the following output:

```
Fixed effects: list(form5)
                      Value  Std.Error     DF   t-value p-value
(Intercept)       0.9494409 0.04454948 242313   21.3121  0.0000
log(LEstimateUSD) 0.9446858 0.00085428 242313 1105.8231  0.0000
Length            0.0008438 0.00008986 242313    9.3908  0.0000
Width            -0.0004550 0.00008287 242313   -5.4904  0.0000
SignedSigned:T   -0.0012238 0.00353482 242313   -0.3462  0.7292
 Correlation:
                  (Intr) l(LEUS Length Width
log(LEstimateUSD) -0.150
Length            -0.013 -0.112
Width             -0.003 -0.115 -0.530
SignedSigned:T    -0.053 -0.084  0.030  0.008

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-14.9512335  -0.6121961  -0.1732241   0.5049592  12.6118409

Number of Observations: 242339
Number of Groups: 22
```

The variables of Lower Estimates, Length, and Width of the paintings are all statistically significant but Signed-TRUE is not, with the baseline of Signed-FALSE. The estimates suggest that the mean sales price for a painting is increased by \$9446.86 with \$1000 USD increase in Lower Estimate. This reinforces the earlier analysis that the lower estimate heavily influences the final sales price. It also suggests that a painting's sales price would increase by \$0.00085 by increase in 10 inches of Length and decrease by \$0.00083 by increase in 10 inches of Width.

Next we scrutinize the random effects. The first step is to examine the intraclass correlation coefficient (ICC), which can be calculated as:

$$\frac{0.2057266}{0.2057266 + 0.5916618}$$

This means that the average correlation of sales price among the paintings within the same auction house is about 25.8%. While it's not close to 1, it does suggest that using a linear mixed model is justified since there is some variation among the auction houses, interpreted as clusters.

Confidence intervals for this model provide more diagnostic insights.

```
Approximate 95% confidence intervals

 Fixed effects:
                        lower          est.         upper
(Intercept)        0.8621250390  0.9494408590  1.0367566790
log(LEstimateUSD)  0.9430114506  0.9446858228  0.9463601951
Length             0.0006677267  0.0008438491  0.0010199715
Width             -0.0006174148 -0.0004549902 -0.0002925657
SignedSigned:T    -0.0081519700 -0.0012238122  0.0057043456
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: SiteID
                    lower      est.     upper
sd((Intercept)) 0.1497539 0.2057266 0.2826199

 Within-group standard error:
    lower      est.     upper
0.5899982 0.5916618 0.5933301
```

This allows us to determine the significance of the random effects. We are 95% confident that the variance component for the intercept lies in (0.14975, 0.28262) and since it does not contain 0, we conclude that the random intercepts are significantly different from 0. Translated, this means that different auction houses exhibit statistically significantly different mean sales prices.

Below is a plot of the EBLUPs of the random effects from this model.

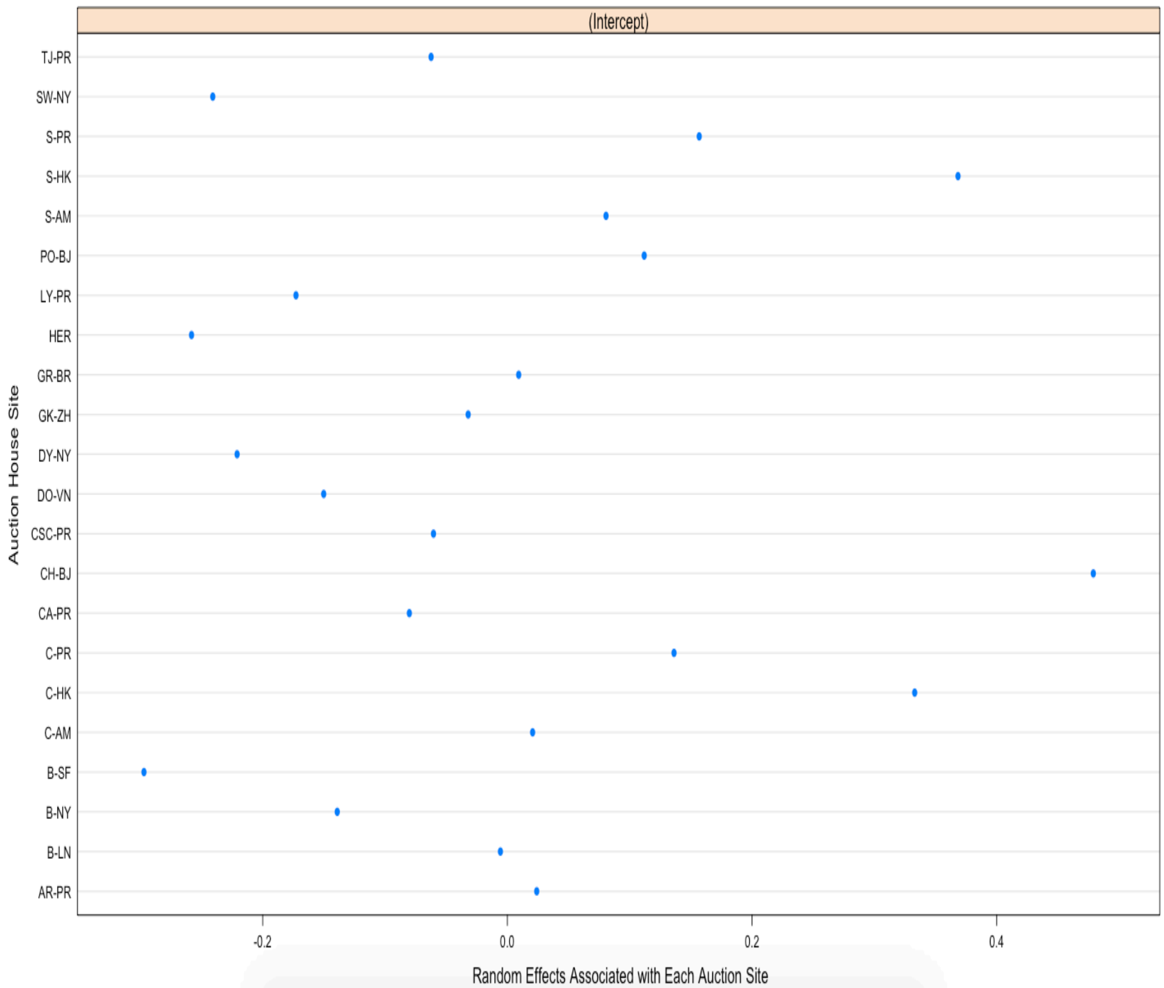

Figure 3A: EBLUPs estimates of random effects

It seems that the random effect for Bonham's-San Francisco has the largest negative value, which corresponds to the outlier mentioned in Figure 2C. These effects estimates seem reasonably dispersed, thus indicating that a choice of auction house does produce different mean prices of paintings. The three largest effects are from China Guardian- Beijing, Christie's Hong Kong, and Sotheby's Hong Kong. For example, an increase $332.98 USD to sales price is attributed to Christie's Hong Kong or more precisely paintings from Christie's Hong Kong would exhibit $ 332.97USD in change of mean sales price for an additional change of 1000 USD. These exemplify the influence of the Chinese art markets and a further examination of this subfield might be worthwhile in future study.

Conclusion

This project examines the ways that a choice of auction house might have an effect on the sold prices of the paintings. The data available is best analyzed by a multilevel model, which naturally leads to the use of both fixed and random effects. We discovered that most of a painting's value derives from its lower estimates given by their respective auction house. Finally, we quantify the extent of the random effects might have on the average means of sales prices and conclude that each auction house exhibits different amounts. With some much data available, there is much more to explore and it is the hope that this project contributes to the existing literature written by other researchers interested in this topic.

References

Carlin, Bradley P., Louis, Thomas A. Bayes and Empirical Bayes Methods for Data Analysis 2nd. 2000.CRC.

Corbeil R.R., Searle, S.R. "Restricted Maximum Likelihood (REML) Estimation of Variance Components in the Mixed Model." *Technometrics* 18. 31-38. 1976.

Chanel, Olivier, Gerard-Varet, Louis Andre, Ginsburgh,Victor. "The Relevance of Hedonic Price Indices: The Case of Paintings." *Journal of Cultural Economics* 20. 1-24. 1996.

Findlay, Michael. The Value of Art: Money, Power, Beauty. 2014. Prestel USA.

Galecki, Andrzej ,Burzykowski, Tomasz, 2013. Linear Mixed-Effects Models Using R- A Step by Step Approach. Springer.

Jiang, Jiming. Asymptotic Analysis of Mixed Effects Models- Theory, Applications, and Open Problems. 2017. CRC.

Pinheiro, Jose C., Bates, Douglas M. Mixed-Effects Models in S and S-PLUS. 2000. Springer.

McNulty, Tom. Art Market Research: A Guide to Methods and Sources 2nd. 2013. McFarland.

Searle, Shayle R., Casella, George, McCulloch, Charles E. Variance Components. 2006. Wiley.

Tango, Toshiro. "Likelihood-based ignorable analysis for missing data" Repeated Measures Design with Generalized Linear Mixed Models for Randomized Controlled Trials. 2017. CRC.

Verbeke, Geert, Molenberghs, Geert. Linear Mixed Models for Longitudinal Data. 2009. Springer.

'' ''. "How Ignorable Is Missing at Random?" Linear Mixed Models in Practice- A SAS Oriented Approach. *Lecture Notes in Statistics*. 1997. Springer.

West, Brady T., Welch, Kathleen B., Galecki, Andrzej T., Linear Mixed Models- A Practical Guide Using Statistical Software 2nd.2015. CRC.