author: "Svetlana Koroteeva, 300432399"

date: 15.06.2022

Sculpture shipping costs

1. Summary

An art exhibitor plans to launch an online portal for art enthusiasts worldwide to collect art with only a click of a button. The exhibitor is interested in determining how different artifact attributes affect shipping cost. They would like to be able to predict the likely cost of an artifact when they acquire it.

After exploratory data analysis, missing data imputation, several models were built for this project. There were investigation linear models with transformation, generalized additive models (GAM), subset selection and ridge regression.

Based on art dataset predictors influence on Cost was investigated and several models for predicting cost was built and compared. Predictions were estimated , better models selected and price prediction example was given in result section.

2. EDA and data imputation

2.1. Dealing with missing data

Art set includes 6500 artifacts for the analysis exercise. The dataset consists of “Customer Id”, “Artist Name”, “Artist Reputation”, “Height”, Width”, “Weight”, “Material”, “Price Of Sculpture” , “Base Shipping Price” “International”, “Express Shipment”, “Installation Included”, “Transport”, “Fragile”, “Customer Information”, “Remote Location” , “Scheduled Date”, “Delivery Date”, “Customer Location”, “Cost” .

Set summary shows that 7 columns contains missing values. Omitting them leaves only half of the set, which is far more then 25%, so data imputation will be the best way On the coloured matrix, table (Fig. 1) and heatmap (Fig.2) we can see that the affected columns are Width, Height, Weight, Artist Reputation, Material, Remote Location, Transport. Missing values spread across the set and clustering in the mentioned columns.

```
          Customer_Id           Artist_Name     Artist_Reputation 
           0.00000000           0.00000000           0.11538462 
               Height                Width               Weight 
           0.05769231           0.08984615           0.09030769 
             Material     Price_Of_Sculpture   Base_Shipping_Price 
           0.11753846           0.00000000           0.00000000 
        International      Express_Shipment Installation_Included 
           0.00000000           0.00000000           0.00000000 
            Transport              Fragile  Customer_Information 
           0.21415385           0.00000000           0.00000000 
      Remote_Location        Scheduled_Date         Delivery_Date 
           0.11861538           0.00000000           0.00000000 
    Customer_Location                 Cost                      
           0.00000000           0.00000000                      
```

Percent of missing values inside of each category shows that transport , location, weight columns contain the most of missing data.

The imputation had been performed via mice (Multivariate imputation by chained equations) package which implements a method to deal with missing data based on Fully Conditional Specification (see methodology).

2.2. Further EDA

From the first look at the predictors we can see that Customer ID column and Customer location does not have repetition, however there is 51 times the same author name encounter. These columns could be considered for deletion.

Date column might influence and for easier calculation it could be converted to date stamps. In addition, 2 other columns produced, which includes years from scheduled and delivered dated and another column added as scheduled and delivered time difference to see, how timeline influences the shipping price.

The scatterplot matrix (Fig. 3) shows possible non-linear relationships between the response variable Cost and predictors Artist Reputation, Hight, Width; transformations of these predictors should be considered. The response variable Cost is skewed, so a transformation should be considered for dealing with non-normality. The matrix plot shows moderate pairwise correlations between predictors weight and price of sculpture, width and height. Multicollinearity should be investigated. The plot also shows that all the date variables and artist reputation have almost 0 correlation with response variables.

As part of the task scheduled and delivery date were investigated, and difference applied to the dataset. Also Years from these columns were extracted to separate column for further investigation. Time Difference has very low correlation to other data.

Non-numerical values are fit in box plot vs Cost. Almost of the box “flatten” to the bottom line and multiple outliers presented. To cut outliers boxplot method was used. On Figure 4 boxplot for the same predictors are paired before and after.

Looking at the Material-Cost price boxplots (Fig.4)- heavy materials (as metal, stone, marble) affect Cost the most. Waterway shipping is slightly cheaper. Fragile items are cheaper in general, however they have outliers. Express shipment, international shipment, remote location, working or wealthy customer status affect cost only slightly.

As result of Exploratory data analysis for the further investigation we can exclude Customer Id, Artist Name, Times Difference. As scheduled and delivery date have correlation one of the parameters could be excluded (and also times difference comes from normal distribution). Data set was saved as art_f

3. Methodology

3.1. Data imputation

Data imputation have been performed via mice based on “fully conditional specification” or “sequential regression multiple imputation”. There are 2 approaches: single imputation based on means and not account to uncertainty of imputation and Compare to single imputation multiple imputation involved creating multiple sets, and not performing well if observed values are not predictable. The second approach involves making several datasets. MICE uses multiple approach however the mice model in joining both ways. MICE operates under the assumption that the missing data are Missing At Random (MAR). The chained equation process includes steps which are repeated for each missing variable and imputation was updated after each cycle.

1.Mean imputation. 2. One “mean” placeholder set back to missing. 3.It regressed to other variables in the regression model. 4. This missing variable replaced with prediction of this regression model

Optimal number of cycles is specified by the researcher. I have set maxit parameter equals 5. Missing value analysis (Figs. 1-3) showed that all missing variables are spread equally across the set. There was not found a reason to exclude some of the predictors, though artist reputation could be excluded as it is correlated only with the cost and price of sculpture, I left it for further investigation.

Concerning other MICE parameters used for imputation PredictorMatrix and Method, default method, which is used in the current imputation method assumes that continuous data are imputed by predictive mean and matrix tells the algorithm which variables predict missingness in which other variables by default based on correlations between variables and the proportion of usable cases.

Complete function taking previous setting and number of set created. I have chosen 2. There are also options ‘long’ and ‘board’ which could be explored.

Imputed data set saved as “art_to_research.csv”

3.2. Linear regression models

For more clear picture numerical predictors fitted to response variable in scatter plots. (Fig. 5). Predictors which demonstrated non-linearity – basic_shipping_price, height and artist reputation.

To start subset selection we are fitting dataset to linear models and implementing the recommendation from EDA paragraph 2. The first fit includes all the predictors and does not have transformations.

$Cost = \beta_0 + \beta_1 * Price\_Of\_Sculpture + \beta_2 * Base\_Shipping\_Price + \beta_3 * International + \beta_4 * Express\_Shipment + \beta_5 * Installation\_Included + \beta_6 * Fragile + \beta_7 * Customer\_Information + \beta_8 * Scheduled\_Date + \beta_9 * Delivery\_Date + \beta_{10} * Artist\_Reputation + \beta_{11} * Height + \beta_{12} * Width + \beta_{13} * Weight + \beta_{14} * Material + \beta_{15} * Remote\_Location + \beta_{16} * Transport + \beta_{17} * Year\_Scheduled + \beta_{18} * Year\_Delivery$

Plots (Fig. 6) show that there is an indication of non-linearity (some curve Residual vs Fitted plot), non-normality (points not on line QQ plot) and non-constant variance when fitted values increase(Scale-Location plot). Shapiro test shows that there is no evidence that residuals come from normal distribution. Breush-Pagan test shows that null hypothesis of homoskedasticity is rejected and heteroskedasticity (non-constant variance) assumed. Variance Inflation factor indicates that all date predictors have severe multicollinearity which may lead to non-reliable coefficients. Therefore they should be deleted or combined to single variable. Figure 7 contains the plot after removing outliers and highly influential observation 4177, 4072 which appeared after on Residuals vs Leverage plot were detected. Note, with a different scale all previous observation became more visible.

For the next 4 fits the following modification had been done fit2 – removed date columns, fit3 – added log transformation for response variable, fit4 – added log transformation for basic shipment price, fit 5 – added log transformation for width, and polynomial transformation to Artist Reputation and Basic Shipment Price.

Assumptions check plots (Fig.8) demonstrate improvements for s non-normality, non-linearity, non-constant variance. To estimate progress R squared is used. Proportion of variance is accounted for by the model gradually increased after each modification and Fit5 looks especially promising.

Fit1	Fit2	Fit3	Fit4	Fit5
0.5341752	0.5465103	0.5671831	0.5765060	0.7177433

3.3. GAM

Generalised additive model could be another approach for multiple linear regression. It provides flexible smoothing functions of some predictor variables, and interest focuses on inference about these smooth functions, parameter that directly controls the smoothness of the curve, or estimated predictive accuracy.

2GAM fits included to the project.

The first fit includes smoothing terms applied to all numerical predictors and log transformation is kept for the response variable. In this model fragility seems to be less significant with p-value >0.05.

For the second GAM model the interaction between Artist Reputation and Width was included.

According to gam check function, which performs residual diagnostic, the first fit might require more degrees of freedom for Price of Sculpture as it has low p-value k almost equal edf and k_index <0. Similarly for Weight. It was changed for the second fit.

	k'	edf	k-index	p-value	
s(Price_Of_Sculpture)	9.00	8.92	0.87	<2e-16	***
s(Base_Shipping_Price)	9.00	8.27	1.01	0.70	
s(Artist_Reputation)	9.00	3.47	1.00	0.36	
s(Height)	9.00	3.07	1.01	0.68	
s(Width)	9.00	2.78	0.99	0.29	
s(Weight)	9.00	1.00	0.97	0.02	*

	k'	edf	k-index	p-value	
s(Price_Of_Sculpture)	19.00	17.70	0.86	<2e-16	***
s(Base_Shipping_Price)	9.00	7.75	0.99	0.320	
s(Artist_Reputation)	9.00	1.00	1.03	0.975	
s(Height)	9.00	3.29	0.99	0.260	
s(Width)	9.00	1.00	1.01	0.690	
s(Weight)	9.00	1.00	0.98	0.065	.
te(Artist_Reputation,Width)	22.00	13.28	0.98	0.145	

Summary table, shows that p-value of s(Weight) is not significant and could be linear for gam1 and in gam 2 smooth function for width and weight could be excluded too.

Approximate significance of smooth terms:

	edf	Ref.df	F	p-value	
s(Price_Of_Sculpture)	8.916	8.998	242.554	<2e-16	***
s(Base_Shipping_Price)	8.266	8.850	367.732	<2e-16	***
s(Artist_Reputation)	3.468	4.314	1821.970	<2e-16	***
s(Height)	3.073	3.865	79.388	<2e-16	***
s(Width)	2.782	3.549	18.473	<2e-16	***
s(Weight)	1.000	1.000	0.232	0.63	

Approximate significance of smooth terms:

	edf	Ref.df	F	p-value	
s(Price_Of_Sculpture)	17.705	18.768	141.049	< 2e-16	***
s(Base_Shipping_Price)	7.747	8.595	357.586	< 2e-16	***
s(Artist_Reputation)	1.000	1.000	18.047	2.23e-05	***
s(Height)	3.288	4.175	64.173	< 2e-16	***
s(Width)	1.000	1.000	0.311	0.577	
s(Weight)	1.000	1.000	0.022	0.882	
te(Artist_Reputation,Width)	13.280	15.569	32.918	< 2e-16	***

Models were assessed with R-squared, Adjusted R-squared and RSE. The results for the second gam fit is slightly higher and these results are slightly better than for linear models from previous paragraph.

Statistic	GAM
RSE	0.2942
R-squared	0.8764
Adj. R-squared	0.8753

Statistic	GAM
RSE	0.2899
R-squared	0.8802
Adj. R-squared	0.8789

Another way to estimate models are Akaike Information Criterion (AIC) and Bayes Information Criterion. The fit5 and both gam models were combined to the table.

According to both BIC and AIC criteria the best model is the second GAM model which included interaction between Artist_Reputation and Width

modname	aicval	bicval
non interation	2605	2892
interation	2036	2439
linear	3533	3784

3.4. Subset Selection

All the models still contain many predictors and it is not very obvious which of them are important for cost prediction. “Subset selection” methods was used for this purpose to see how they could be reduced. For the forward selection the empty model is taken and then new predictors added to it, opposite for backward, where all predictors were added on the first instance and model is being estimated for this subset selection with Cp, AIC, BIC, or adjusted R2 , which is approximately related to the test MSE. The smallest value of the test MSE achieves the optimum balance between the bias-variance trade-off.

Library leaps contains “best subset”, “forward” and “backward” functions which could produce different set of predictors. For the first selection the number of values was set to 8 and 1 best model for each sum number of predictors. Figure 9 shows that the best model have been chosen model with predictors and they are “Price_Of_Sculpture","Base_Shipping_Price,"Artist_Reputation","Height" and “Express Shipment”.

(Intercept)	Price_Of_Sculpture	Base_Shipping_Price	Artist_Reputation
-416.186817	1.405734	12.133146	826.256676
Height	Express_ShipmentYes		
7.849482	84.003047		

Forward and Backward selection left the same 5 predictors.

3.5. Ridge regression model

An alternative to subset selection methods ridge regression with cross validation was applied as one of the shrinkage methods od statistical learning with glmnet method. The features fit to the model were "Price_Of_Sculpture","Installation_Included","Fragile", "Base_Shipping_Price","Artist_Reputation","Height","Width","Weight","Material","Transport","Customer_Information","Express_Shipment","International","Remote_Location" the same as for subset selection.
Best lambda =38. (Fig. 10). R-squared of predictions variables was calculated and equal 0.68. Which means that only 0.68 percent of data could be explained by data.

All the predictors are more then 0, however it is roughly seen that the more significant (with bigger coefficients) are the same “Price of sculpture”, “Base_Shipping_Price”, “Height”, “Express Shipment” . However “Width” coefficient is in the same group. Next follow “Material”,”Transport”,”Customer_Information”, ”Installation”

```
                                          s1
(Intercept)                       -4.037218e+02
(Intercept)                                .
Price_Of_Sculpture                 1.256177e+00
Installation_IncludedYes           4.855014e+01
FragileYes                        -2.419367e+01
Base_Shipping_Price                9.869553e+00
Artist_Reputation                  7.801057e+02
Height                             8.082309e+00
Width                              8.253325e+00
Weight                            -9.093315e-06
MaterialBrass                      1.124170e+02
MaterialBronze                     7.805303e+01
MaterialClay                       5.883851e+01
MaterialMarble                     1.910909e+02
MaterialStone                      2.157272e+02
MaterialWood                      -6.404593e+01
TransportRoadways                 -4.902406e+01
TransportWaterways                -7.467261e+01
Customer_InformationWorking Class -3.910036e+01
Express_ShipmentYes                4.413337e+01
InternationalYes                   2.911145e+01
```

3.6. Fits with reduced features

Fitting reduced set of predictors to linear and gam model, gave the following results. It seems that BIC and AIC value higher than for full models. Also linear model and gam with 5 predictors are estimated as better rather than with 11 for both AIC and BIC criterions. It seems that the rough cut off based on lower or higher coefficients when they are extremely small did not work well for model selection. So for the next comparison from this step first linear model with 5 predictors is taken, and gam model with 5 predictors.

modname	aicval	bicval	r_sq
linear 5	3554	3707	0.813
linear 11	6474	6640	0.813
gam 5	2623	2791	0.848
gam 11	6000	6197	0.848

3.7. Prediction estimation

For these step 5 models were selected based on previous research.

1. fit5 <- lm(log(Cost)~ Price_Of_Sculpture + poly(Base_Shipping_Price,10)+International+Express_Shipment+Installation_Included+Fragile+Customer_Information+poly(Artist_Reputation,8)+Height+log(Width)+Weight+Material+Remote_Location+Transport, data=X_train)
2. gam2 <- gam(log(Cost)~ s(Price_Of_Sculpture, k=20) + s(Base_Shipping_Price)+International+Express_Shipment+Installation_Included+Fragile+Customer_Information+s(Artist_Reputation)+s(Height)+Width+Weight+Material+Remote_Location+Transport+te(Artist_Reputation,Width), data=X_train)
3. fit.new1 <-lm(log(Cost)~ Price_Of_Sculpture + poly(Base_Shipping_Price,10) + International + poly(Artist_Reputation,8)+Height+Express_Shipment, data=X_train)

4. gam.new1 <-gam(log(Cost)~ s(Price_Of_Sculpture) + s(Base_Shipping_Price) + s(Height)+s(Artist_Reputation)+Express_Shipment, data=X_train)
5. Ridge regression with the following predictors "Price_Of_Sculpture","Installation_Included","Fragile", "Base_Shipping_Price","Artist_Reputation","Height","Width","Weight","Material","Transport","Customer_Information","Express_Shipment","International","Remote_Location"

Before predicting y hat data, the existing set should be divided into training and test in proportion 20/80 and refit again.

I would like to highlight 2 important steps in the process of prediction and prediction estimating

- All predictions had been done for log(cost) , so for transforming them back exp() function was used over all y hat set
- To bring all the models to the same estimators MSE and R-squared had been used.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \qquad R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

All 5 models had been resulting in a merged table. Best results showed GAM model with all main predictors. Next reduced GAM model. Ridge model showed not high MSE however the lowest % of data explaned by model. And it is opposite for both linear models, the r-squared were better for linear models, however they show much more significant mean squared error.

names	mses	r_sqs
FIT	582	0.724
GAM	71.6	0.776
FIT5	692	0.704
GAM5	79.9	0.75
RIDGE	92.8	0.695

4. Results

For results 2 models had been selected.

GAM model with 13 predictors and interaction.:

Price_Of_Sculpture,Base_Shipping_Price,International,Express_Shipment, Installation_Included, Fragile, Customer_Information, Artist_Reputation, Height, Width, Weight, Material, Remote_Location, Transport, interaction (Artist_Reputatio*Width):

Smooth function applied for 4 predictors and interaction. 77.6% of data could be explained by the model.

and GAM model with 5 predictors:

Price_Of_Sculpture,Base_Shipping_Price,International,Express_Shipment, Artist_Reputation and Height

Smooth function applied to all the predictors. 75% of data could be explained by this model.

Smooth terms are making explaining of individual predictor influence more complicated.

Confidence interval, which reflects the uncertainty around the mean predictions was calculated by presented next formulas.

```
upr <- exp(p$fit + (2 * p$se.fit))
lwr <- exp(p$fit - (2 * p$se.fit))
```

There is an example table produced by code from the last section. For this example 11-th row from test data was selected.

conf_names	mod1	mod2	real
lwr	1316	1423	-
fit	1421	1510	1506.1
upr	1535	1603	-

This means that if Price of Sculpture=193, No installation, Not Frigile, Base_Shipping_Price =80.23, Artist_Reputation = 0.59, Height = 16, Width = 6, Weight = 61911, Material =storn, Transport= Roadways, Customer_Information=Working Class, No Express_Shipment, International, Not remote location item will be shipped that with 95% of confidence the price with the lay between 1316 and 1535 for first model and between 1423 and 1603 for second model.

5. Next steps

What could be done next? The predictability of the final models is not very high and not all the abalable options were investigated in this project. There was only one example with interactions and it was successful. More models could be built with different combinations of different interacting predictors. There was only one method of selecting valuable predictors covered and on Shrinkage method – Ridge regression. Lasso could be next trial, as well as other tools for selecting predictors. Smooth term degrees of freedom changing was not fully investigated.

Also the code could be rewritten to be able quickly switch parameters and produce and access different models in more automotive way.

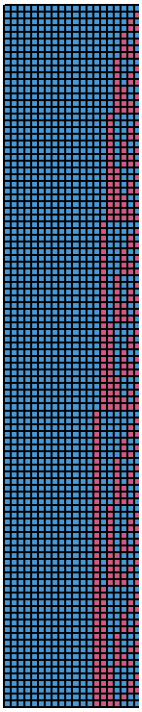
Concerning interpretation of current model it would be good to add how change for each particular predictor influence cost changes and investigate how to reverse smooth terms.

To keep less complexity of the model, liner models should be chosen. Complexity is the week point of the selected models. This might be investigated with special methods.

6. Reference

- Multiple imputation by chained equations: what is it and how does it work? - PMC (nih.gov) – MICE methods information
- https://data.library.virginia.edu/getting-started-with-multiple-imputation-in-r/ -missing data investigation with mice.
- Video tutorial: https://www.youtube.com/watch?v=fsqAi9ddHU0 – data imputation
- "An Introduction to Statistical Learning with Application R" Gareth James, Daniela Witten, Trevor Hastie, Robert
- Course materials
- Ridge Regression in R (Step-by-Step) (statology.org) – r-squared
- Predict in R: Model Predictions and Confidence Intervals - Articles - STHDA -prediction interval and confidence intervals,
- r - Confidence interval for GAM model - Cross Validated (stackexchange.com) -confidence interval for GAM
- autosmooth.pdf (anu.edu.au) -GAM

1. Appendix



	Artist_Reputation	Material	Remote_Location	Transport	
3362	1	1	1	1	0
500	1	1	1	0	1
249	1	1	0	1	1
111	1	1	0	0	2
213	1	0	1	1	1
130	1	0	1	0	2
43	1	0	0	1	2
21	1	0	0	0	3
220	0	1	1	1	1
122	0	1	1	0	2
55	0	1	0	1	2
23	0	1	0	0	3
59	0	0	1	1	2
20	0	0	1	0	3
10	0	0	0	1	3

Figure 1. Missing values pattern matrix plot and numbers (fragment)

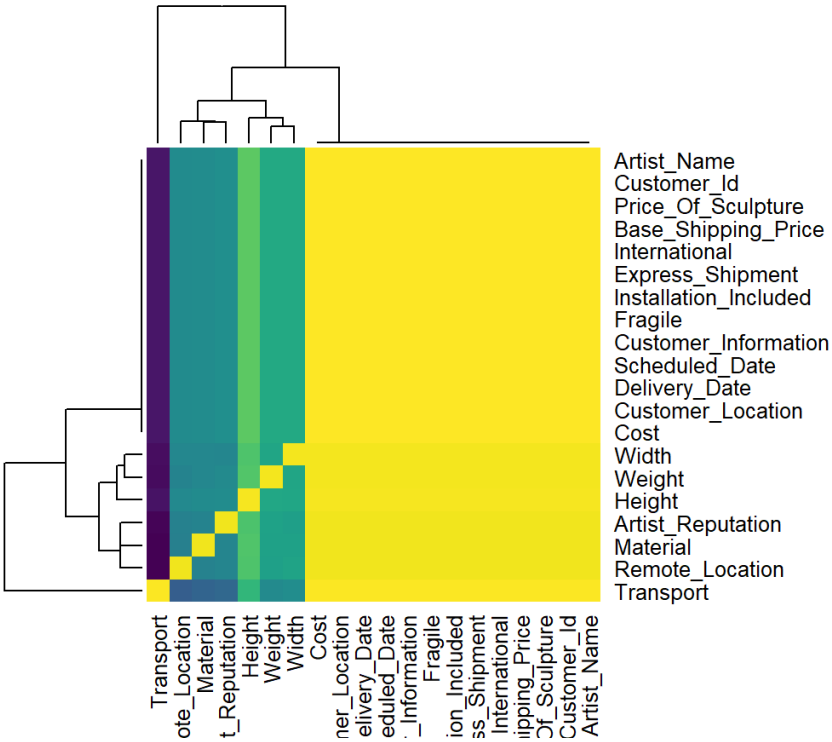


Figure 2. Missing values heatmap

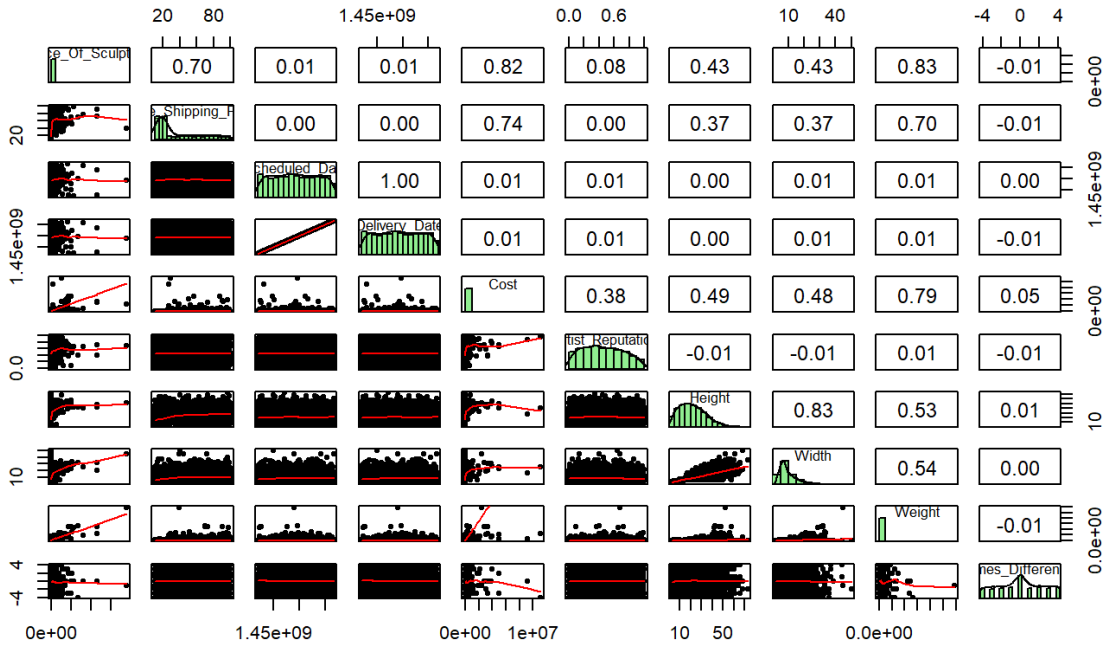
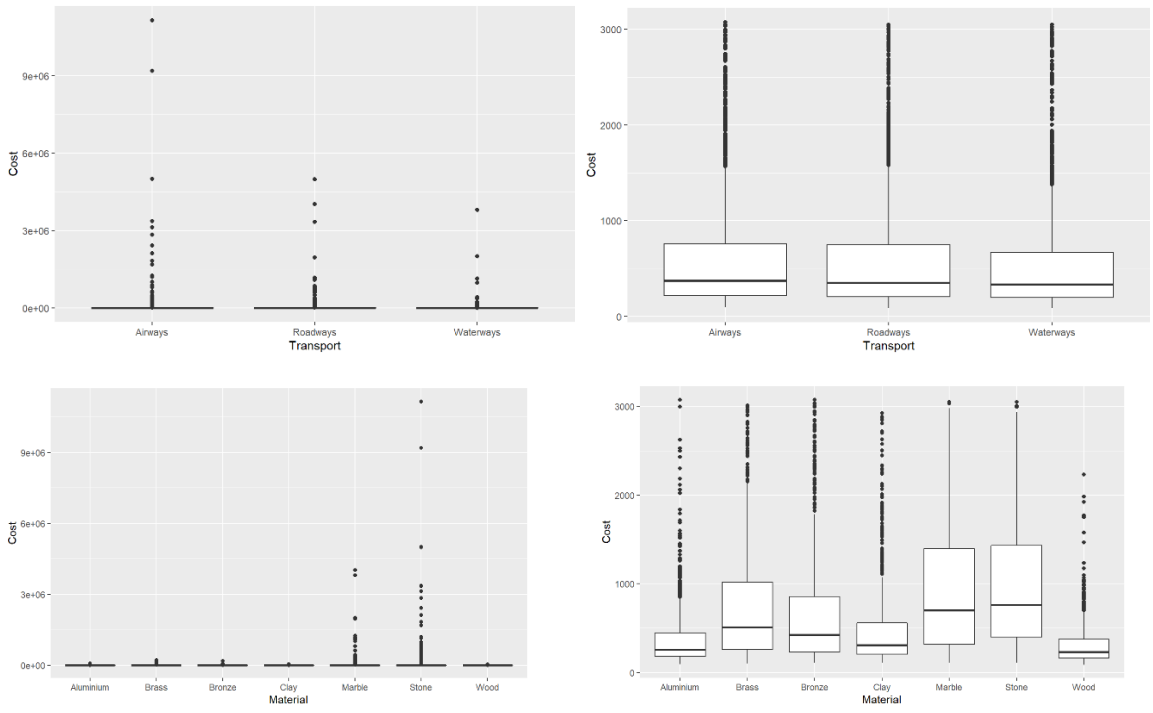


Figure 3. Data matrix plot



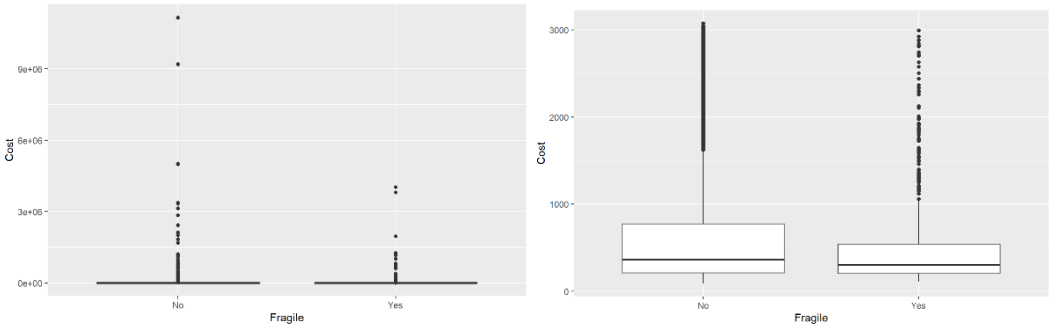


Figure 4. Boxplots for non-numeric predictors

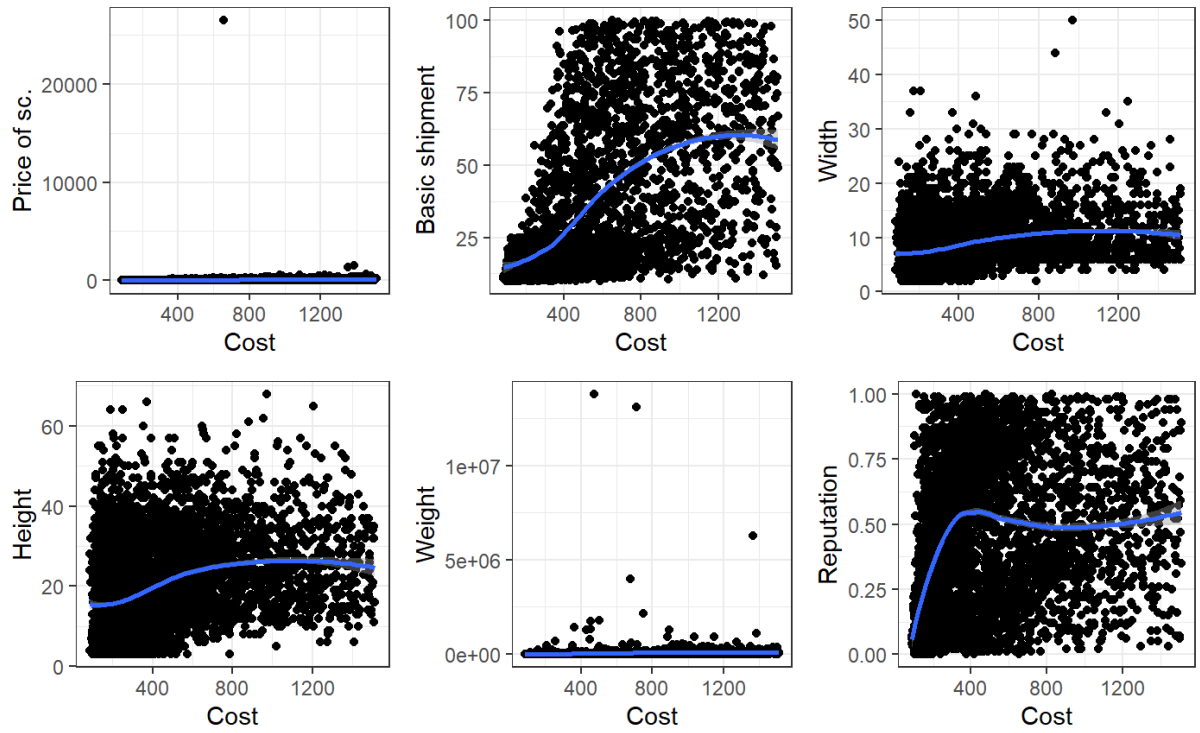
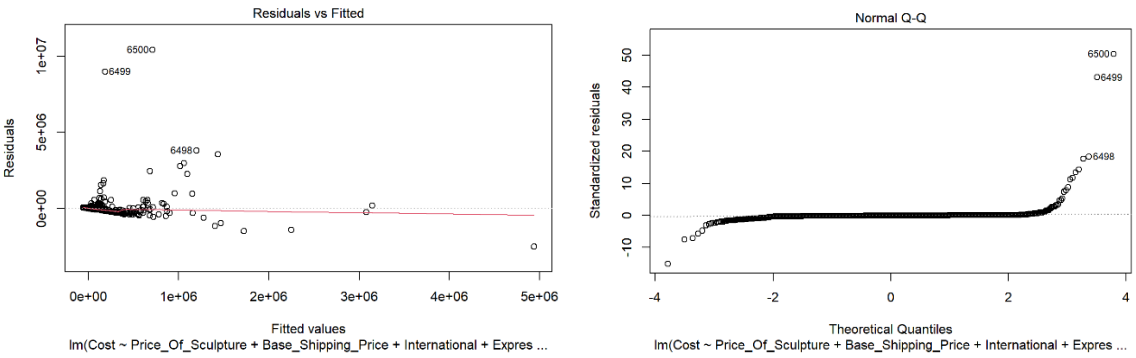


Figure 5. Scatter plots for response variable vs numerical predictors.



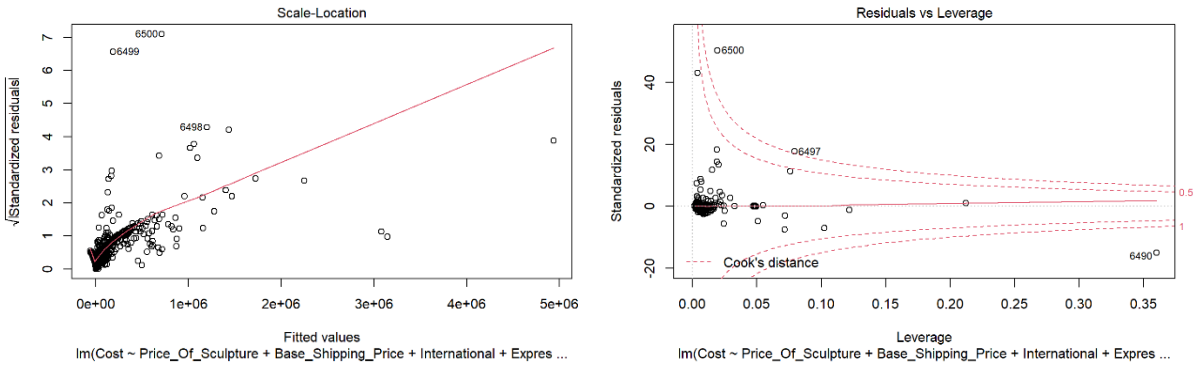


Figure 6. The initial linear model assumptions check plots

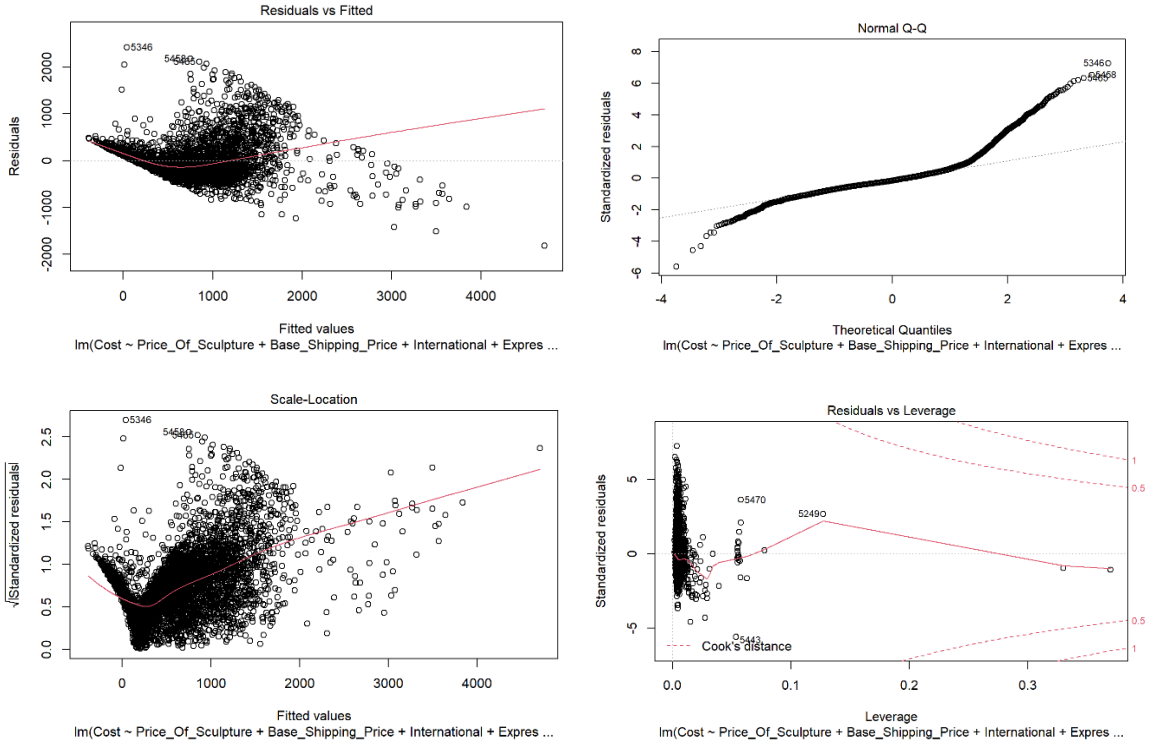
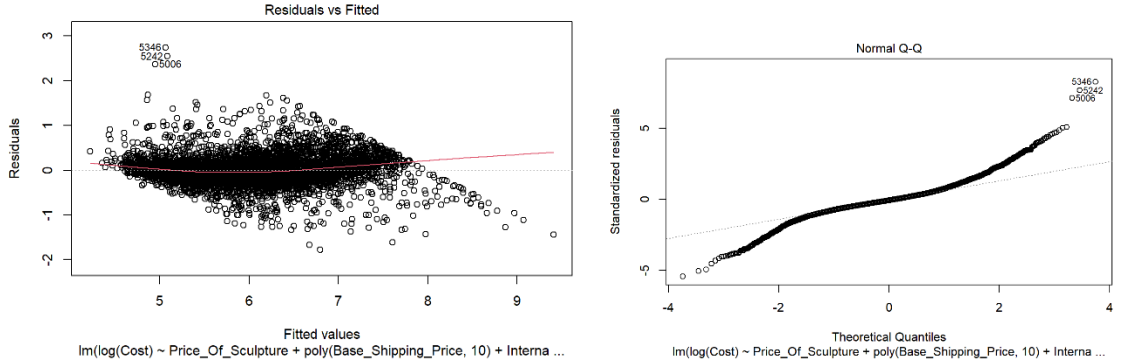


Figure 7. Linear model assumptions check without outliers and highly influential points



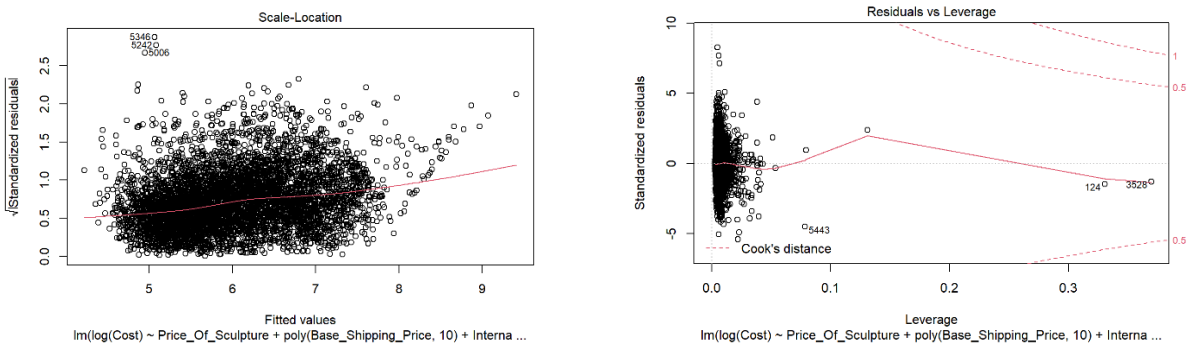


Figure 8. Modified linear model assumptions check plot

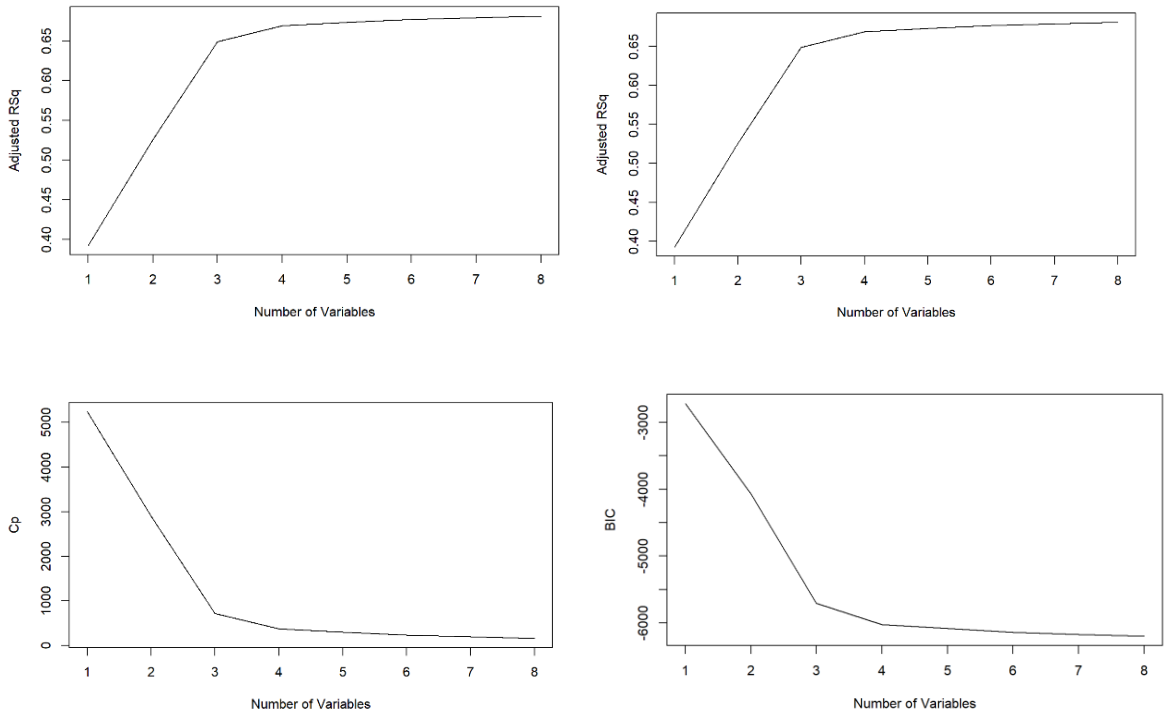


Figure 9. Best subset selection results

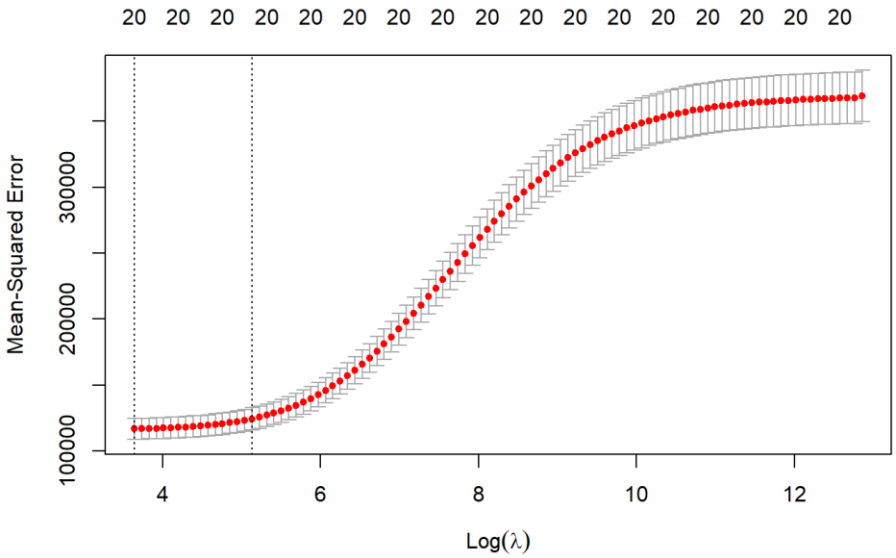


Figure 10. Best lambda selection