Kurtis Hausman

Professor Nandhini Gulasingam

DSC 323

11 November 2024

Road Construction Report

[Appendix:](#_4210cli9ltm5)

For this project, I decided to explore the Road Construction data set, which held 279 total observations. This dataset examined how different independent predictor variables impact the price of a contract bid on a road construction project by the lowest bidder. Road construction contracts in the state of Florida are awarded on the basis of competitive, sealed bids. The contractor who submits the lowest bid price wins the rights to the contract. In the 1980s, the Office of the Florida Attorney General (FLAG) suspected numerous contractors of practicing bid collusion, meaning they would work together to collectively set the winning bid price above the fair or competitive price in order to increase their profit margin. By comparing the bid prices and other important variables [(A1)](#umuh2r5pufzu) of the rigged contracts to the competitively bid contracts, FLAG was able to establish invaluable benchmarks for detecting future bid-rigging. FLAG collected data for 279 road construction contracts. This data includes our independent x-variables representing different information about a particular road project. These will be used in a linear regression model to predict our independent y-variable, the price of a contract bid on a road construction project by the lowest bidder. The goal of this project is to be able to fit a linear regression model by selecting the combination of predictors and data points that can most accurately predict the contract bid price.

To begin this project I imported the given dataset into R for preliminary inspection. I noticed that there were a number of NA values in the “Len” (Length of road project in miles) column of the dataset [(A2)](#s8z9dnvmbgc9). To handle these I decided to get rid of observations that had NA values, because I did not want to attempt to fill them and incorporate my own bias, for choosing which method to impute the missing values, as I lacked domain knowledge. After removing the observations which had NA values I had 217 observations remaining. I saved this dataset, road\_construction\_cleaned.csv and used that for my SAS analysis.

After Importing the dataset into R, I created dummy variables for Districts 1 through 4 to represent the categorical “dist” variable in a format suitable for regression analysis [(B1)](#gyxstrr6i3w9), 1 would represent the project being in that district, and 0 would represent the project not being in that district. An observation being in the fifth district is implicitly represented by the absence of a 1 in all of the four dummy variables, allowing the model to use District 5 as a reference category. This approach helps avoid multicollinearity issues in the model.

Next I began to explore the data with plots to get an initial understanding of it. The histogram [(B2)](#inn3gnu6zuc8) shows extreme right skew, which led me to anticipate I would need to log transform my dependent variable: contract bid price.

The scatterplot matrix [(B3)](#b3xamea4xbqr) showed only two predictors that had any linear relationships with the dependent variable: Day (Estimated number of days to complete work) and Fair (DOT Estimate of a fair contract price) positive linear trends.

The boxplots of each categorical predictor: Contract Status, Subcontractor Utilization, and District. The boxplots of price by status [(B4)](#hck4lkb53l2x) and price by sub [(B5)](#efc0tl8tsl9) did not show very much insight, but price by district [(B6)](#ylr65r1kquad) showed that road construction projects in district 4 tend to have much higher contract prices, on average, than the other districts.

The frequency plots [(B7)](#n274v09md83j) help us to understand this further. The frequency by Subcontractor Utilization suggests that subcontractor use is relatively rare in this dataset. The Frequency by Status of Contract , indicates that competitive bidding is more common, but there is a notable portion of fixed contracts, which might be of interest in investigating bid collusion or irregularities. Frequency by Districts (1-5) tells us that most contracts are concentrated in District 1 and District 5 , while Districts 2, 3, and 4 have significantly fewer contracts.

The descriptives for the full model [(B8)](#f2jn6h9ilj86) told us a bit more about the data. Our Adj-R^2 being 0.9811 and R^2 being 0.9826 tell us that about 98% of our outcome in y (price) can be explained by using our predictors (x-variables) in the model. However, we can see from the p-values all being greater than 0.05 that all of our predictors in this model are insignificant. The high RMSE of 247.98985 suggests that, on average, the model’s predictions are about 247.99 units away from the actual values.

The residual plots help us confirm that suspicion that we need to transform our dependent variable. The Normal Cumulative Distribution vs. CDF of Studentized Residual plot [(B9)](#udhdajdtgnl) shows a significant S-shape bend. It does not support normality and linearity throughout, suggesting that the residuals deviate noticeably from a normal distribution.

The Studentized Residual vs. Predicted Value model [(B10)](#ew0d357kt0tt) has concerns too. There seems to be a tight cluster around the left side which makes it difficult to see any trend. The points do not show Independence (constant variance) and are not randomly scattered. Overall this residual model does not help to improve our confidence in our current model, so I intend to act on the assumption that I will need to transform my y-variable.

After running the full model, I found a few outlier points, but for this project I chose to do my training set and test set split prior as well as model selection prior to removing any points. Due to only having 217 observations and a lack of domain knowledge, I wanted to be hesitant about removing points. I only removed the outliers on the train set, and left the test set alone. This is so that when I tested my data, it would be given a fair assessment, as the test data would most closely resemble what you would receive in a real-world production environment.

I then decided to make my log transformation on the dependent variable, price, and rename the new variable ln\_price [(C1)](#hscq1tlok9ai). Following this, I went through all of the data exploration again. I began to explore the data with plots to get an initial understanding of it.

The histogram [(C2)](#fhi5mzsk42te) now shows a normal distribution, which is my first indicator that the log transformation was a good tactic to use.

The scatterplot matrix [(C3)](#k51fim9lcobf) now did not show any predictors with a linear relationship, I am guessing this is because of the amount of predictors still impacting this model. The only two predictors that had any linear relationships with the dependent variable in the previous plot: Day (Estimated number of days to complete work) and Fair (DOT Estimate of a fair contract price) now showed log trends, further representation of the log transformation.

The boxplots [(C4)](#eqdpvjydn7ht) did not change drastically from this transformation, but allowed me to see that the ln\_price by sub was on average higher for contracts that did utilize a subcontractor. The frequency plots are the same.

The descriptives for the log transformed full model [(C5)](#bg4mgjjgcvpx) told us a bit more about the data. We see that only 8 of the 17 predictors are now insignificant based on their p-values being greater than 0.05: district1, district3, bid, pla, pb, pe, pm, sub. This leads me to think that when model selection takes place later on some of these will not likely be in our final model. Our Adj-R^2 being 0.7972 tells us that about 79% of our outcome in y (ln\_price) can be explained by using our significant predictors, penalizing for the insignificant ones. R^2 being 0.8132 tells us that about 81% of our outcome in y (ln\_price) can be explained by using all of our predictors in the model. The RMSE of 0.612222 is now much smaller and suggests that, on average, the model’s predictions are about 0.612222 units away from the actual values. These seem much more reasonable for a preliminary model with room for improvement.

The Normal Cumulative Distribution vs. CDF of Studentized Residual plot [(C6)](#pgy229j6waum) shows a normal trend. It supports normality and linearity through its positive linear trend and only curves slightly with no significant S shape bends.

The Studentized Residual vs. Predicted Value model [(C7)](#oyefmphl4l0f) is decent. I would argue that for the amount of data I have, the points do show Independence (constant variance) and are randomly scattered. There is only one point outside of the -3 to 3 y-range, indicating a potential outlier.

The residuals by predictor [(C8)](#g5ww3prvkplp) are important to include now because we can actually view a trend in them from the log transformation. There seems to be more of a cluster around the left of the data, but if you restrict to endpoints just slightly, the points do show Independence (constant variance) and are randomly scattered.

Overall these residual models do help to improve our confidence in our final model. There is no striking pattern indicating problems in the regression analysis. There is no striking pattern indicating problems in the regression analysis

Again there were a few outliers that were flagged but I have decided not to remove any points until after I do my training set and test set split prior as well as model selection.

I used a 75/25 split, 75% of data points for the training set to build my model on and 25% for the testing set to evaluate my model’s performance. I created a dataset named road\_trainingSet [(D)](#qzf8rf8htr2) that held all of the observations in the training set and worked with that from this point forward. I also created a new dependent variable for values only in my training set called train\_y that represents ln\_price for the training set. This ensures data integrity and prevents data leakage during the model training and evaluation process. This also allows for more flexible testing and cross validation as well as easy comparison between the actual and predicted values.

Once I split the data, prior to model selection I needed to get an idea of the potential multicollinearity between my predictors. My Pearson Correlation Coefficient [(E1)](#37x64bkl2wuz) table will allow me to begin to check for potential multicollinearity among the predictors. Since none of the predictors have a correlation value greater than the absolute value of 0.9, this test passes.

Next I checked the (Variance Inflation Factor) VIF for each predictor [(E2)](#h45nidjdofe8). Since none of the predictors have a VIF greater than 10 we can’t assume any collinearity among the predictors. Both of these two results indicate that the possibility of multicollinearity being an issue in our model is low.

Following this I went into the model selection process to build a final model from the training set. I used three selection methods and compared them all in order to select the best possible model.

My first selection method, Adj-R^2 [(F1)](#bi2q1xja7b),gave me 13 predictors: fair, ratio, status, district1, district2, district4, day, len, pla, pe, pm, ptc, sub. It had an Adj-R^2 of 0.8209 indicating that 82% of our outcome in y can be represented by our significant predictors, with penalty for insignificance and an R^2 of 0.8353 indicating that 83% of our outcome in train\_y can be represented by our predictors. However, this selection method did not give the p-values of its predictors, so some of them could be insignificant.

My second selection method, Stepwise [(F2)](#jcbjveec9wbc), gave me 11 predictors: fair, ratio, status, district4, bid, day, len, pe, pm, ptc, sub. It had an R^2 of 0.8321 indicating that 83% of our outcome in train\_y can be represented by our predictors. However, this method gave 2 insignificant predictors: ptc(0.0760), bid(0.12840).

My third selection method, Backward [(F3)](#r8dnvuaoqra5) gave me 9 predictors: fair, ratio, status, district4, day, len, pla, pm, ptc. It had an R^2 of 0.8275 indicating that 82% of our outcome in train\_y can be represented by our predictors. In addition, this method confirmed that all of its predictors were significant, because they all had p-values less than 0.05.

For this reason I chose the model that was given by the Backward selection process because it has a comparable R^2 value to the other two, while using the least amount of predictors and we know that all of those predictors are significant.

After running my final model [(G1)](#8esi49l8d99u), selected from the backward process, on the training set. I was able to get new model descriptives.

My Adj-R^2 being 0.8174 tells us that about 81% of our outcome in train\_y can be explained by using our significant predictors. The RMSE of 0.57518 is now smaller and suggests that, on average, the model’s predictions are about 0.57518 units away from the actual values. This RMSE is lower and Adj-R^2 is higher than the full model. Indicating that we have produced a good final model from selection.

In addition to those, our F-Value for our final model is 81.55 which is quite large with a p-value less than 0.001. From this we can reject the null hypothesis, concluding that at least one variable is significantly associated with the response variable. The model explains a significant amount of variance in the dependent variable (train\_y).

Using the Pearson Correlation Coefficients [(G2)](#ym49yxecwdbt) and the VIF values [(G3)](#35txpwxyfivt) for the predictors in the final model, we can again conclude that multicollinearity among predictors is not an issue for our model.

While checking for Outliers and Influential points on the train set, I found one point, observation number 5 [(G4)](#sasmdg5uzco3). At this point I allowed myself to remove data points as I am preparing my model to have the greatest predictive power that it can for the test set.

After removing observation 5, I checked to see if my model metrics improved [(H1)](#ufexqzsp1kaj). My Adj-R^2 improved to 0.8248 which tells us that about 82% of our outcome in train\_y can be explained by using our significant predictors, penalizing for significance.

My R^2 improved to 0.8346 which tells us that about 83% of our outcome in train\_y can be explained by using our predictors. The difference between R^2 (0.8346) and Adj-R^2 (0.8248) is small, indicating that the model is not overly complex and doesn’t include many unnecessary predictors.

The RMSE of 0.56380 is now even smaller and suggests that, on average, the model’s predictions are about 0.56380 units away from the actual values. This RMSE is lower and Adj-R^2 is higher than the final model that included observation 5, indicating that we have further improved our model.

In addition to those, our F-Value is now 85.20 is quite large with a p-value less than 0.001. From this we can reject the null hypothesis, concluding that at least one variable is significantly associated with the response variable. The model explains a significant amount of variance in the dependent variable (train\_y).

Now that I had the final model, I analyzed my residuals and confirmed my assumptions about my improved final model.

The Normal Cumulative Distribution vs. CDF of Studentized Residual plot [(H2)](#pncgl6wf2kv) shows a normal trend. It supports normality and linearity through its positive linear trend and only curves slightly with no significant S shape bends

The Studentized Residual vs. Predicted Value model [(H3)](#b63b1yo1n6ka) looks very similar to before. Again, I would argue that for the amount of data points I have, the points do show Independence (constant variance) and are randomly scattered. There is no longer a point outside of the -3 to 3 y-range, leading me to believe that the point could have been observation number 5 that I removed.

The majority of the residuals by predictor [(H4)](#k5acpz2nx6tc) have one small issue There seems to be more of a cluster around the left of the data, but if you restrict to endpoints just slightly, the points do show Independence (constant variance) and are randomly scattered.

Overall these residual models do help to improve our confidence in our final model, but are not much different from the residual plots for the full model. The fact that residuals did not change much indicates that removing predictors did not significantly impact the model’s ability to fit the data. This is positive, as it means the removed predictors were likely contributing little to the model’s predictive power.

Once I had my final model built from the training set, I was ready to test its predictive power on my test set. The phat value [(I1)](#vnt8qsrr1enw), represented the predicted y-value (ln\_price) that my model predicted for the observations in the test set. I then examined the metrics to evaluate how the model had performed.

The Root MSE (Standard Error) of 0.76897 [(I2)](#o2btq4247n9) indicates that, on average, the magnitude of prediction errors (with more weight on larger errors) is about 0.76897 units. The MAE (Mean Absolute Error) of 0.59230 [(I2)](#o2btq4247n9) indicates that, on average, 0.5923 units from actual values, with all errors treated equally. Since RMSE is higher than MAE, it suggests that there are some larger errors in the predictions that are pushing up the RMSE.

The R^2 is 0.72262 [(I1)](#vnt8qsrr1enw), taken from yhat^2. This indicates that 72.26% of our output in the dependent variable (ln\_price) is explained by the predictors.

The Adj-R^2 is 0.6659 [(I3)](#vc4b1h2zznf), indicating that 66.59% of our output in the dependent variable (ln\_price) is explained by our significant predictors, with penalty for overfitting.

The CV R^2 = 0.112 [(I4)](#odv4hrv8e58e). This is the absolute value of the difference between our Test R^2 and the Train R^2. Since it is below 0.3, we can determine we are not overfitting the training data and our model has good generalizability to explain new data

Because the RMSE, R^2, and ADJ-R^2 are lower for the training set we can say that the model is better fit for the training set than on the testing set [(I5)](#xngb3x6vbxoj). This makes sense because the model was fit on the training set and makes predictions on the test set. In addition, since the CV-R^2 is below 0.3, we are not overfitting the training data and our model can be generalized to explain new data which is a crucial aspect for creating a good model.

To further explore the predictive power of the model further I split the entire dataset into multiple subsets (or "folds") and used them iteratively for training and testing. I used stepwise as the selection method for this cross validation, with 25 percent of observations saved for the test set.

The Average standard Error (ASE) for the train set was 0.37503 [(J1)](#wjjlql0qlk5). The model fits the training data closely, meaning that on average, the model’s predictions differ from the actual values in the training set by a squared error of 0.3750.

The Average standard Error (ASE) for the test set was 0.53613 [(J1)](#wjjlql0qlk5). The model’s error on the test set is a bit higher, indicating it generalizes fairly well but is slightly less accurate on unseen data.

The difference between these values is 0.1611. Ideally you want the difference between ASE values on the train and test set to be extremely small as this indicates that the model has a similar level of error on both sets, suggesting that it generalizes well without overfitting or underfitting.

This difference between our ASE values for our train and test sets suggests some degree of overfitting, but not extreme. The model’s generalizability is fairly solid.

The ASE plots help us to see how the ASE for both the train and test sets changed as more predictors were removed in order to visualize how the optimal ASE was achieved [(J2)](#z2pcuef1a3bx).

Since I transformed our y-variable at the beginning of this analysis, in order to get a proper model equation I must retransform it. The original equation is for the natural log of the dependent price variable, represented as ln\_price. After retransforming, using the equation included in my output [(K)](#gvon2f3ry5um), we will get coefficients that we can interpret directly to represent the original dependent variable, price (of a contract bid by the lowest bidder) [(K)](#gvon2f3ry5um) .

The intercept (4.42686), is the baseline price of the contract bid ($) when all other variables are zero. While it may not have a direct interpretation in real terms, it serves as the starting point for the model.

fair (0.014223) means that for each additional dollar in the DOT engineer's estimate of fair contract price, the expected price bid increases by approximately 1.42%. This is the predictor with the least significant effect on price.

ratio (90.4082) means that for each one-unit increase in the ratio of the winning bid price to the DOT engineer’s estimate of fair price, the price bid is expected to increase by $90.41. This variable has a substantial impact on the bid price, indicating that as the bid price approaches or exceeds the fair price estimate, the bid price increases significantly.

status (55.32197): means that if the contract is a fixed contract, the price bid is expected to increase by approximately $55.32.This suggests that fixed contracts are generally associated with higher bid prices.

district4 (130.4537) means that if the project is located in district 4 the price bid is expected to increase by $130.45. This is the predictor with the most significant effect on price. This indicates a significant difference in bid pricing based on whether or not the project falls within district 4, which is insightful for preliminary predictions on future data collection.

day (0.5083) means that for each additional estimated day to complete the project, the price bid is expected to increase by approximately $0.51 This small increase suggests that the time required has a minimal, though positive, impact on bid price.

len (7.1168): means that for each additional mile of road project length, the price bid is expected to increase by $7.12. This indicates that longer projects tend to have higher bid prices, as expected due to the increased scope of work.

pla (-0.6121): means that for each additional percentage of costs allocated to liquid asphalt, the price bid decreases by $0.61. This negative coefficient suggests that higher allocations to liquid asphalt might be associated with lower overall bid prices, potentially due to cost efficiencies in materials.

pm (-3.5427): means that for each additional percentage of costs allocated to mobilization, the price bid decreases by $3.54. This negative impact indicates that as more of the budget is allocated to mobilization, the overall bid price tends to decrease, possibly reflecting streamlined logistics or cost savings.

ptc (-2.6191) means that for each additional percentage of costs allocated to traffic control, the price bid decreases by $2.62. This suggests that higher allocations to traffic control are associated with slightly lower overall bid prices, which might reflect efficient planning or specific project requirements.

For the final test of my model’s predictive power I examined how it performed on a specific new point, using values I selected for each of the predictor categories. To do this I analyzed the ranges of confidence and prediction intervals for the value [(L)](#avxp089etjma). Again, the intervals that output are in log-transformed form, representing the natural logarithm of the predicted price in dollars.

The predicted price in log form is 6.8571. Converting this to the original dollar scale by applying the transformation e^6.8571, we get an estimated price of approximately $950.61 (Predicted Price). This predicted price represents the most likely contract bid given the values of the predictor variables for the particular observation.

The 95% Confidence Interval for the mean prediction is from 6.5919 to 7.1224 in log form. Converting these values with the same e^x transformation we get a Lower bound of 729.16 and an Upper bound of 1239.42

This means we can say with 95% confidence that the average contract bid for similar road construction projects with these characteristics is between approximately $729.16 and $1239.42

The 95% Prediction Interval is from 5.7121 to 8.0022 in log form. Converting these values we get a Lower bound of 302.51 and an Upper bound of 2987.52

This mean we are 95% confident that the predicted contract bid for an individual road construction project with similar characteristics will fall between approximately $302.51 and $2987.52

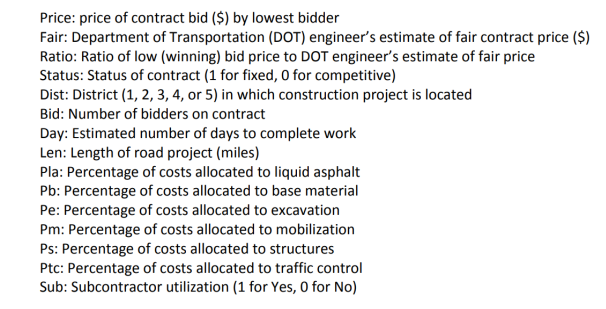
In conclusion, our model built on the training set proved it was capable of delivering predictions with good accuracy on its own test set and also proved that it was capable of generalizing to new datasets as well, increasing its predictive power and reliability.

Throughout this project I recognize there are a few things I recognize I could have done differently with this dataset. First, for further confidence in my model, I could have created multiple models using the models that were given by the other model selection methods. I could have then compared the predictive performance metrics of these different models. This would allow me to add another layer of confidence in the model that performed the best, as I would know for certain it was superior to other models built from the same data set.

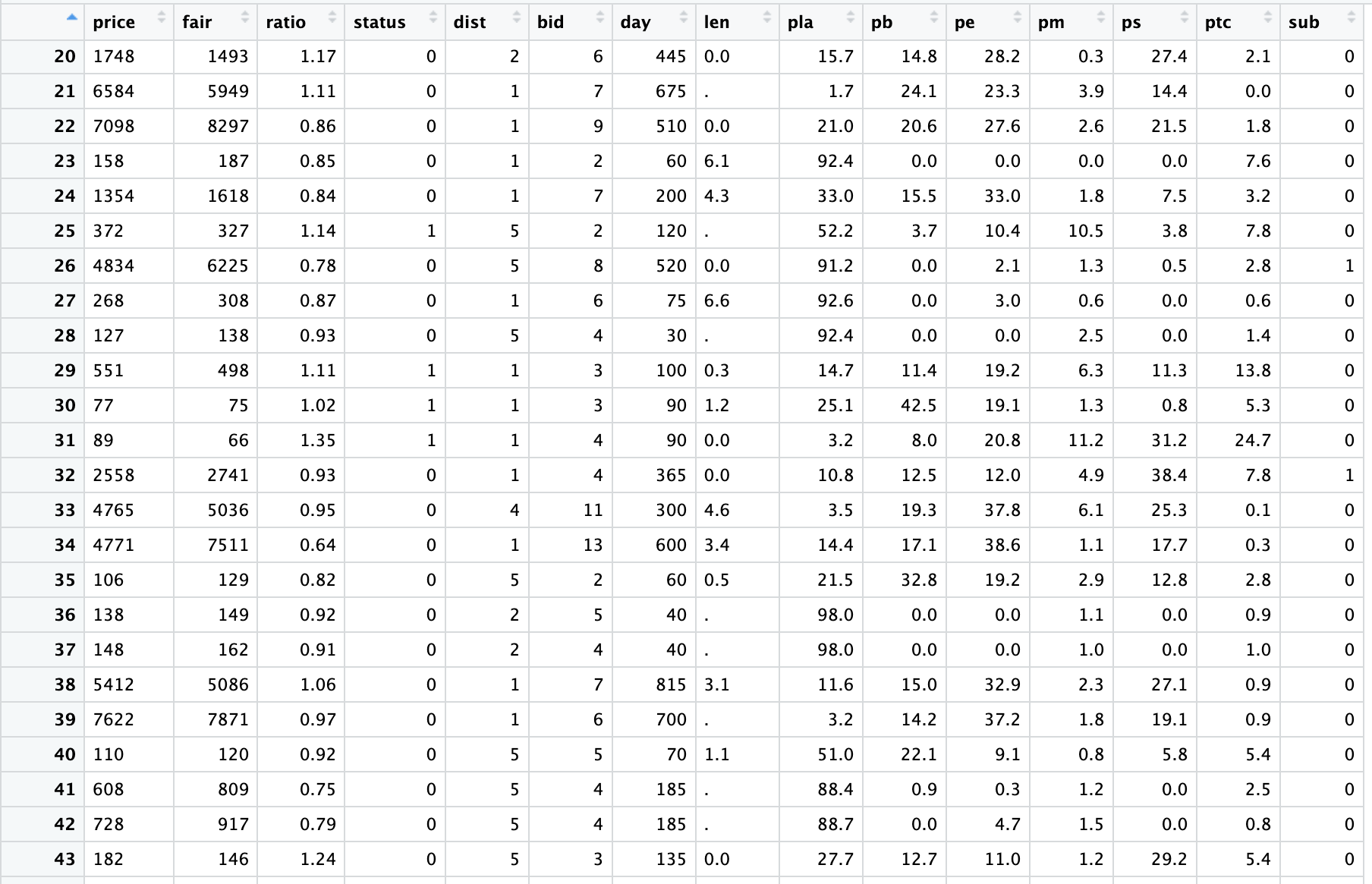
Another thing I could have done differently was to exclude outlier points earlier in the process. If I excluded them after the transformation and prior to splitting into train and test sets and the model selection process, maybe my model would have higher descriptive and performance metrics. However, I do stand by reasoning to not do so. Since this was a relatively small dataset, I knew that a single datapoint had a stronger influence than it maybe would in another dataset. Since I was not the one who collected this data and did not have domain expertise in this field, I did not think I was best suited to be removing many data points. Additionally, by keeping these data points, my model was able to practice predictions on data that most closely resembled a real-world environment, which increases my confidence that this model would perform well on a real-world dataset.

# Appendix:

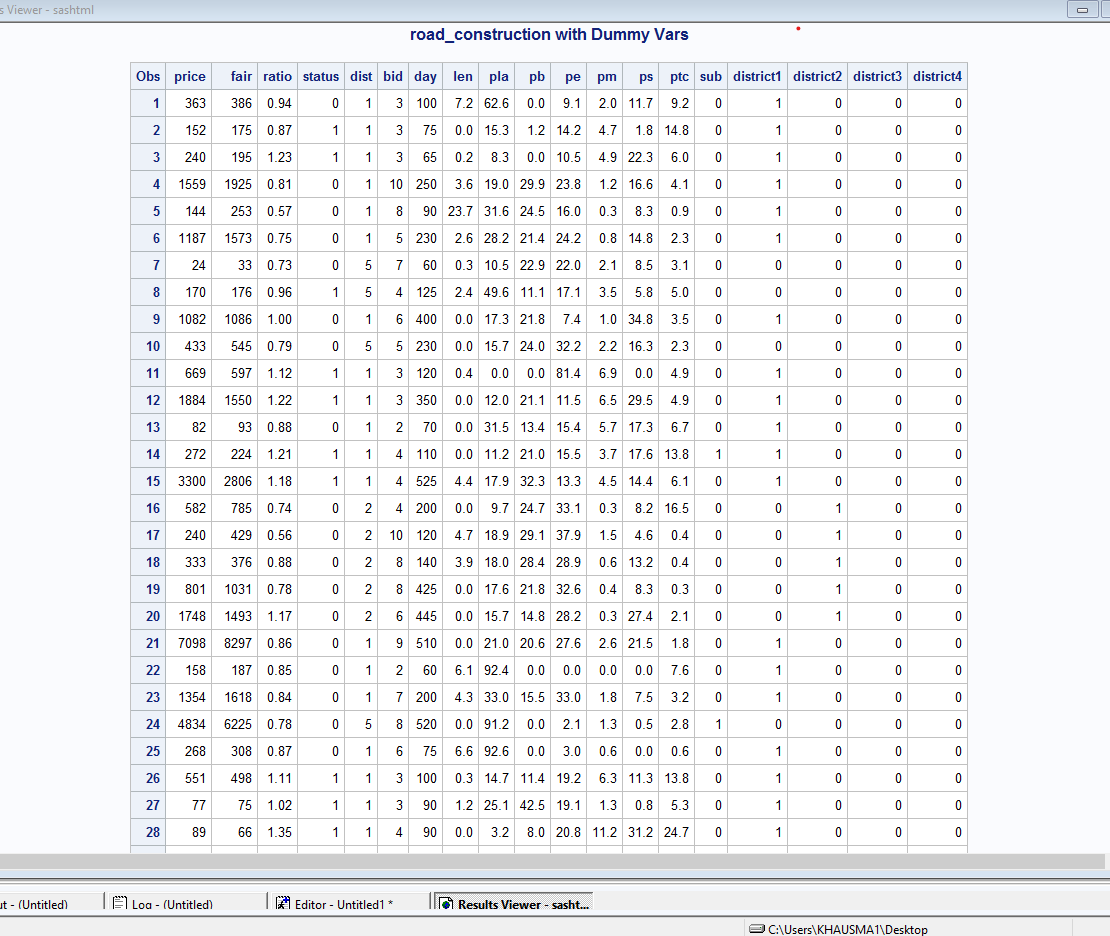
Item A1:



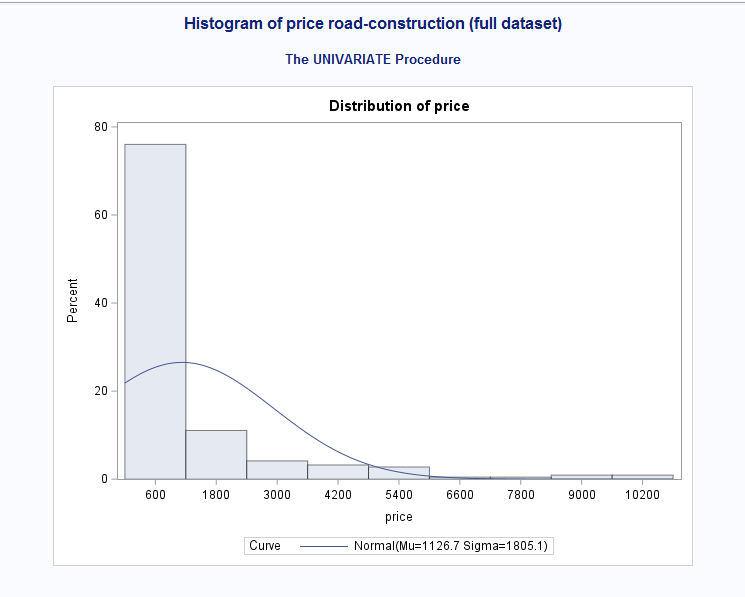
Item A2:



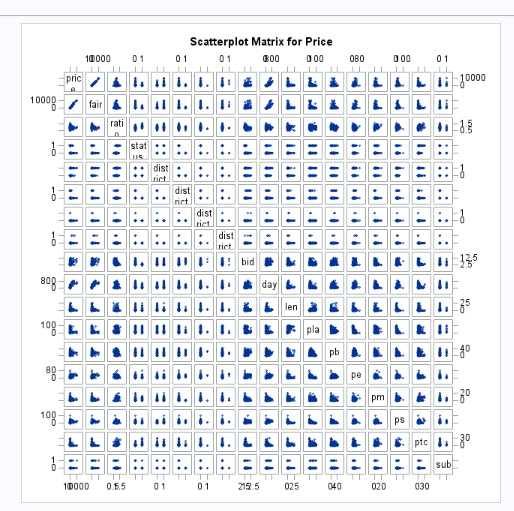
Item B1:



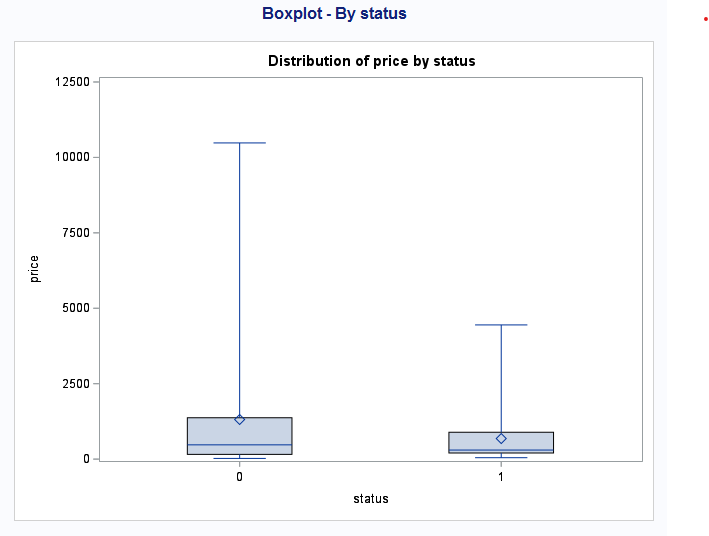
Item B2:



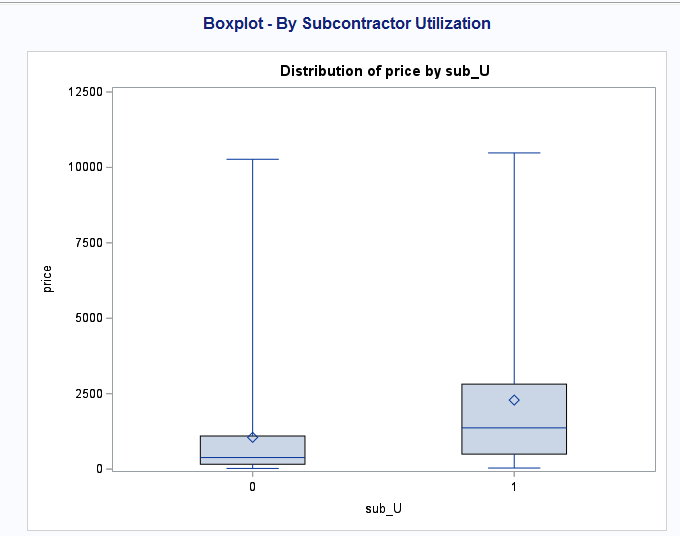
Item B3:



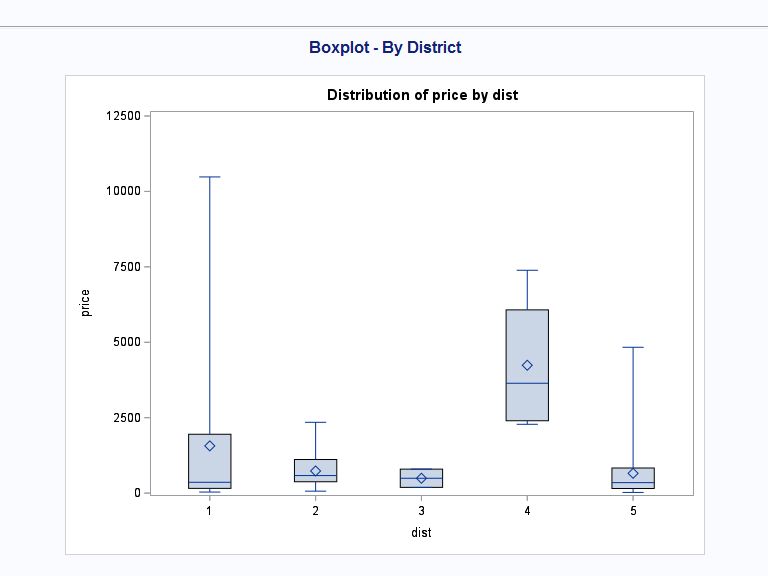
Item B4:



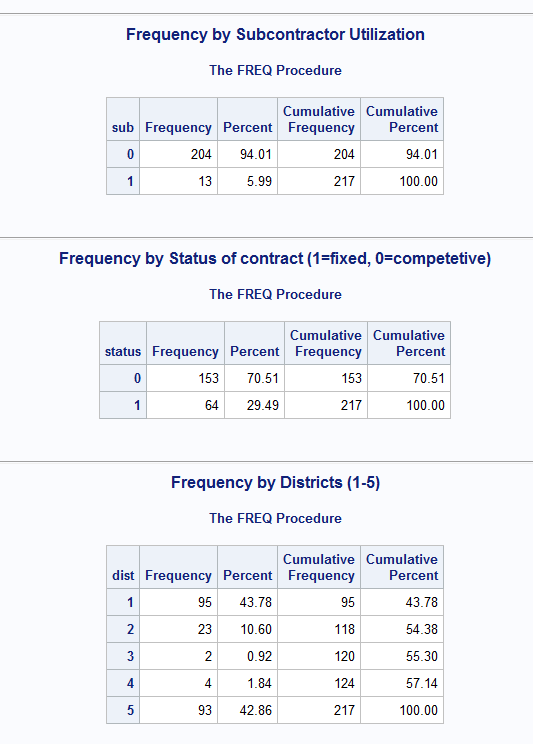
Item B5:



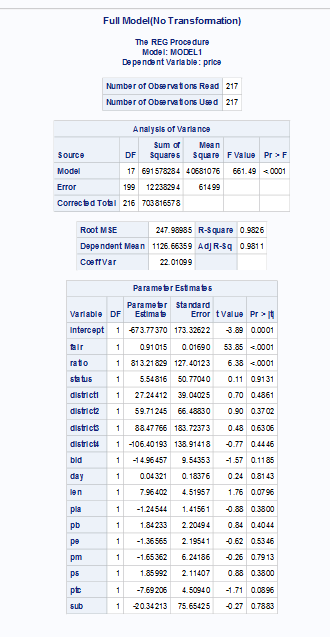
Item B6:



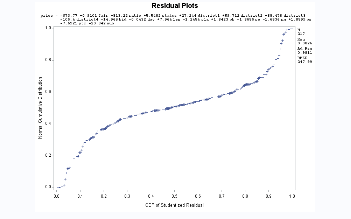
Item B7:



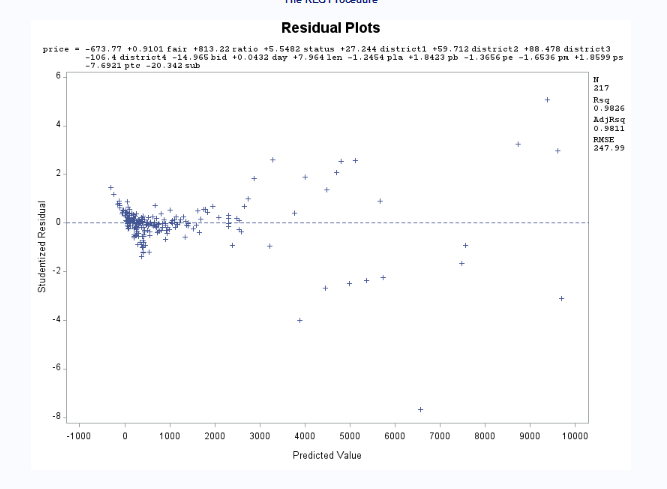
Item B8:



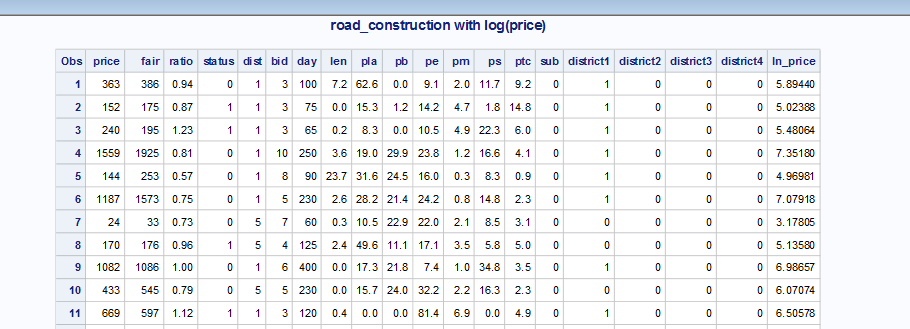
Item B9:



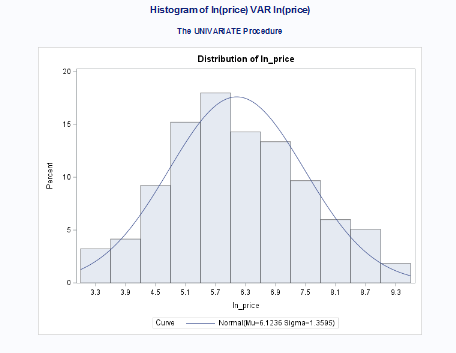
Item B10:



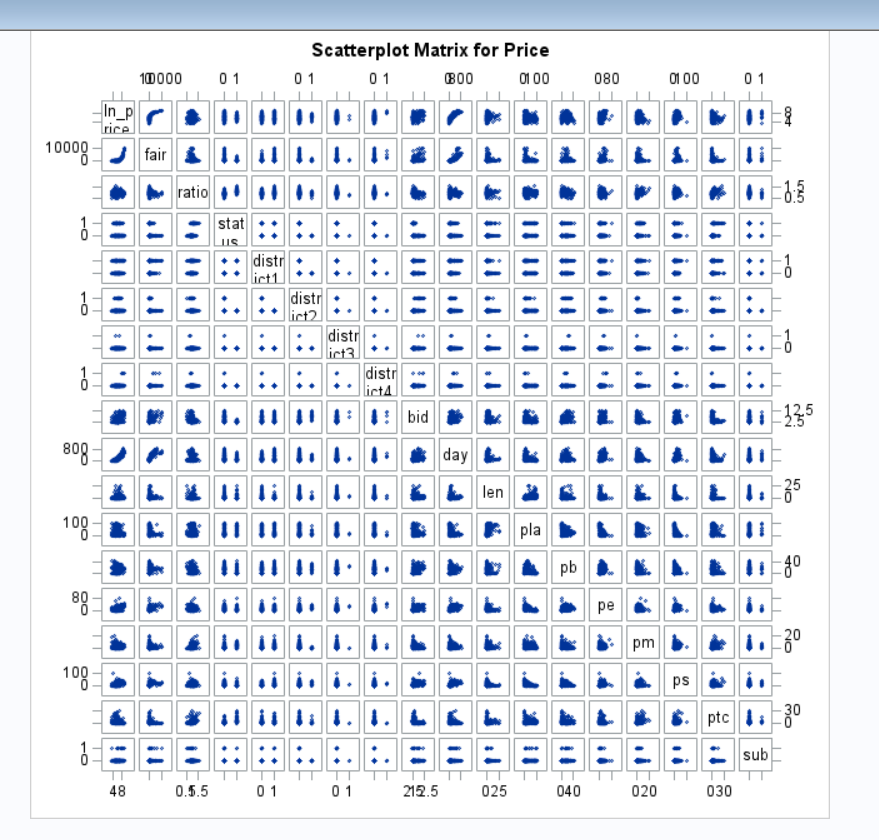
Item C1:



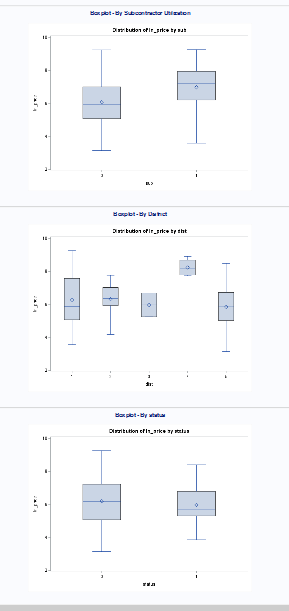
Item C2:



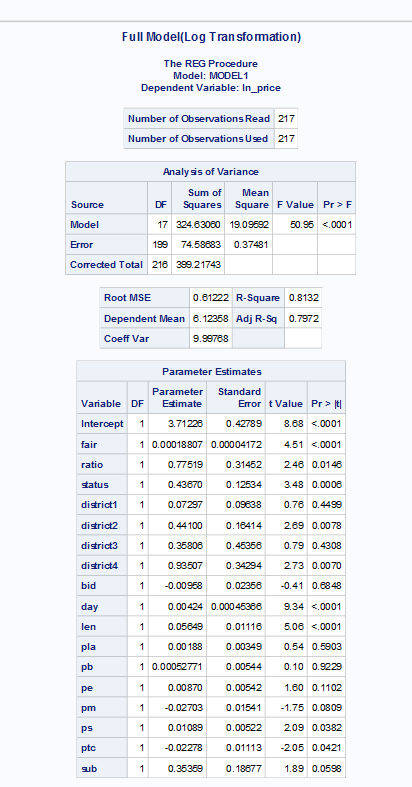
Item C3:



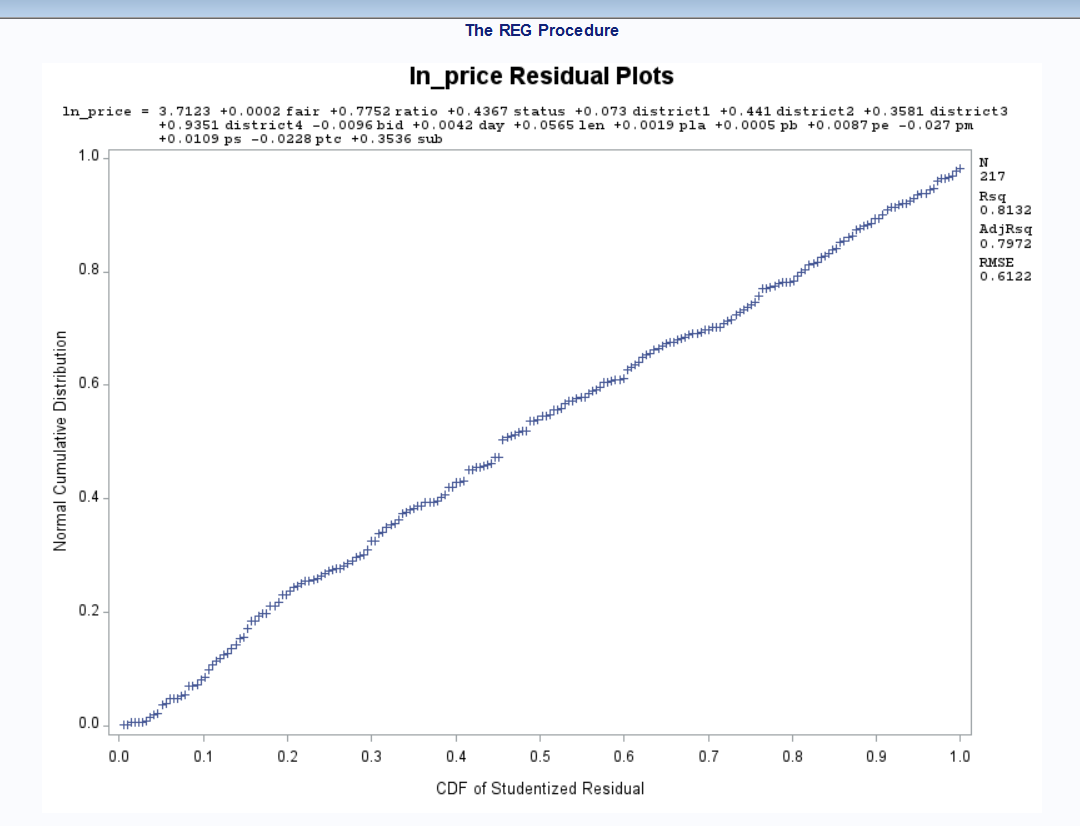
Item C4:



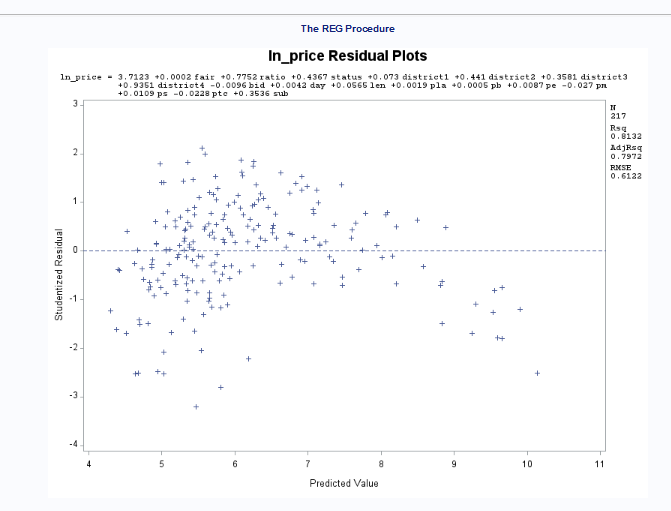
Item C5:



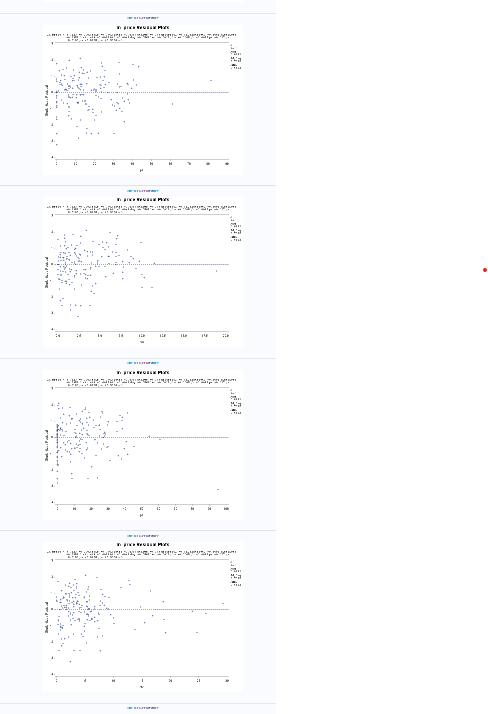
Item C6:

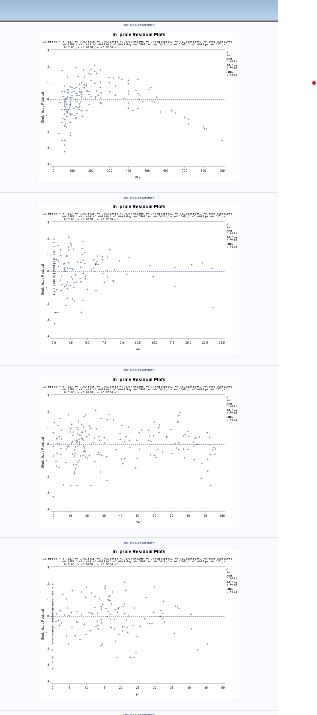


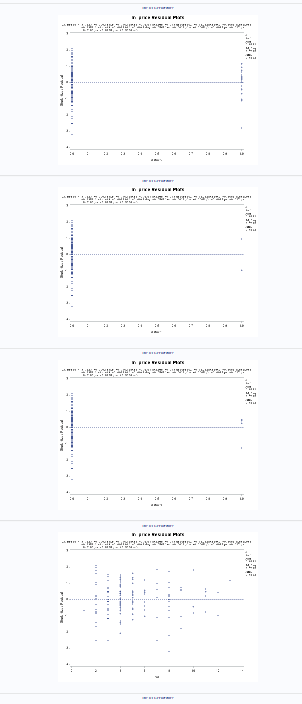
Item C7:



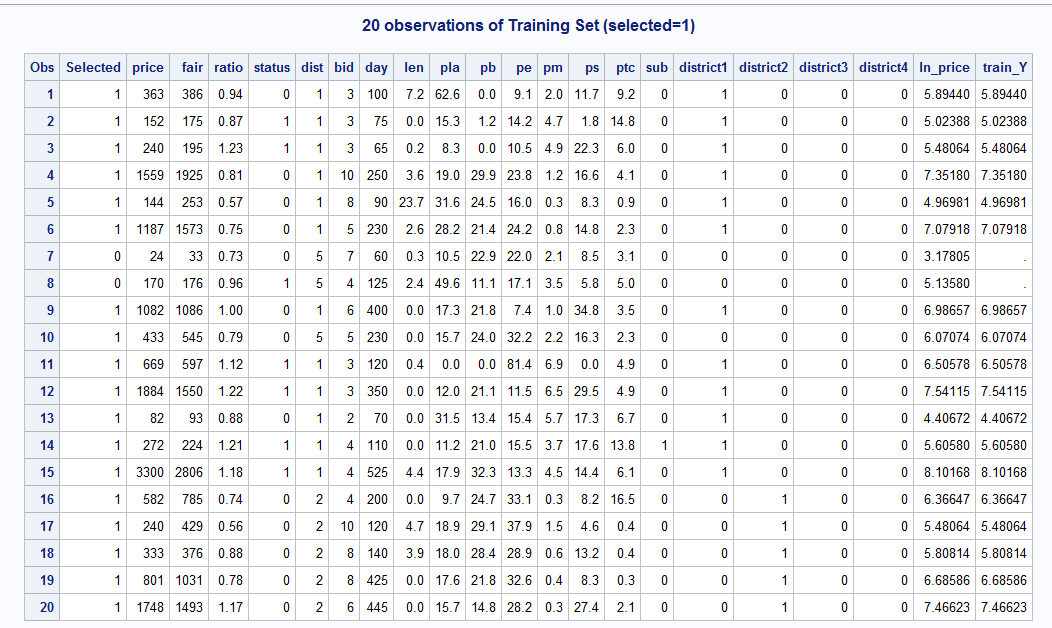
Item C8:



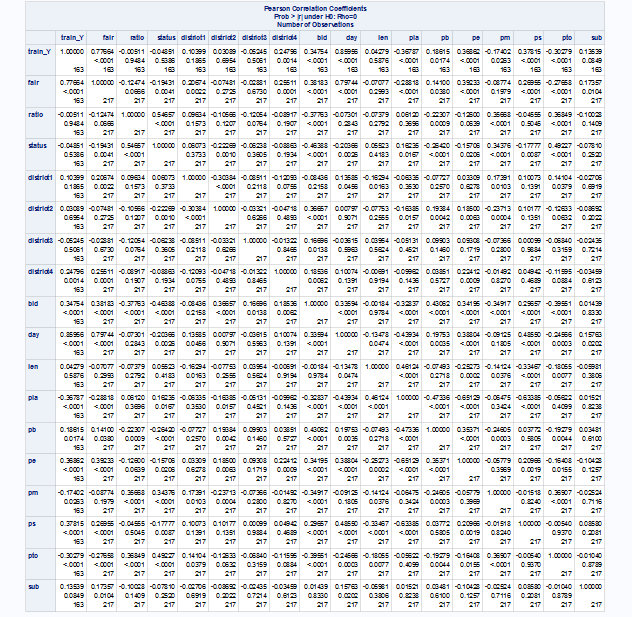




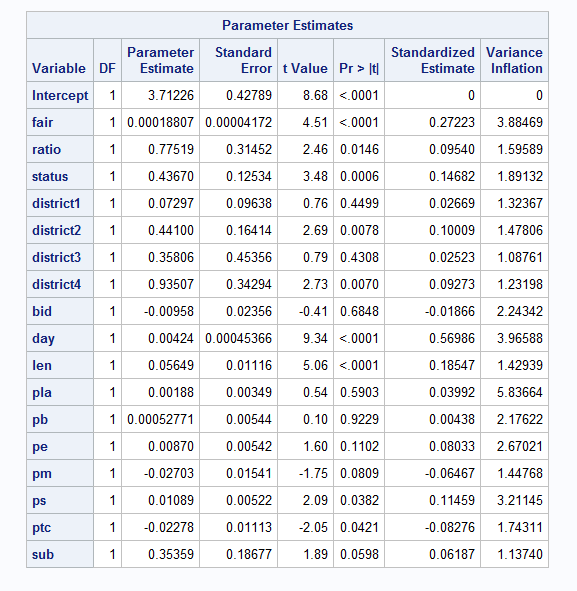
Item D:



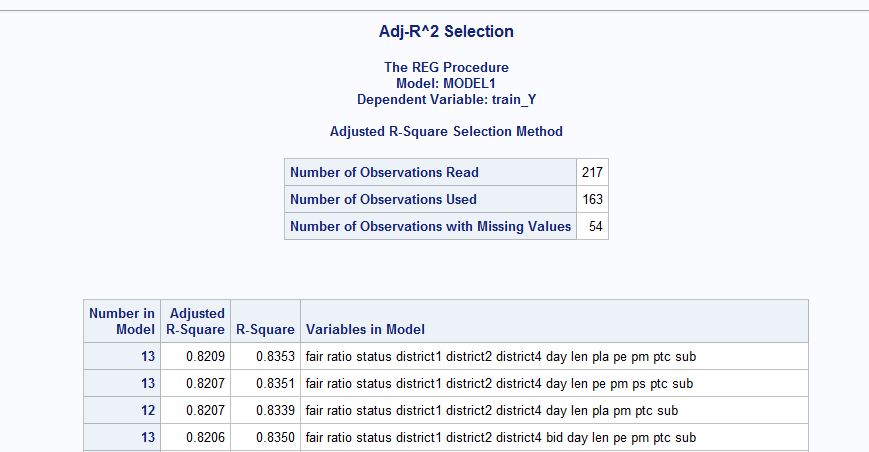
Item E1:



Item E2:



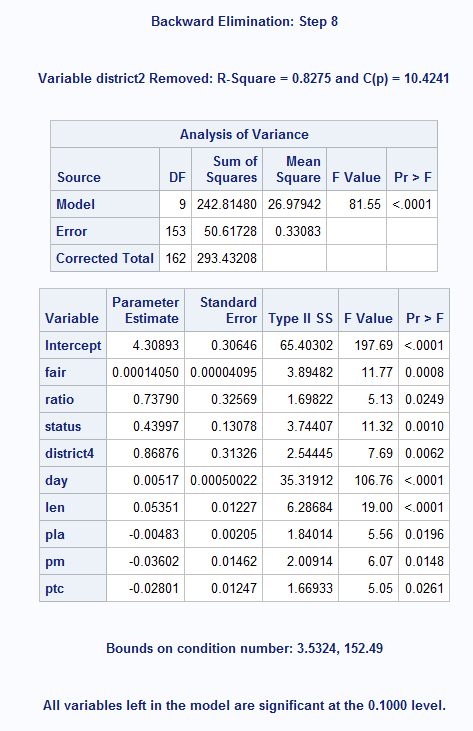
Item F1



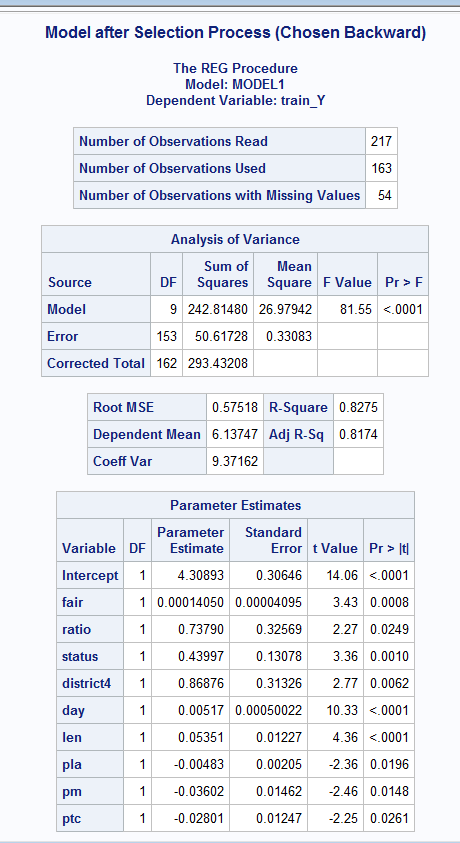
Item F2:



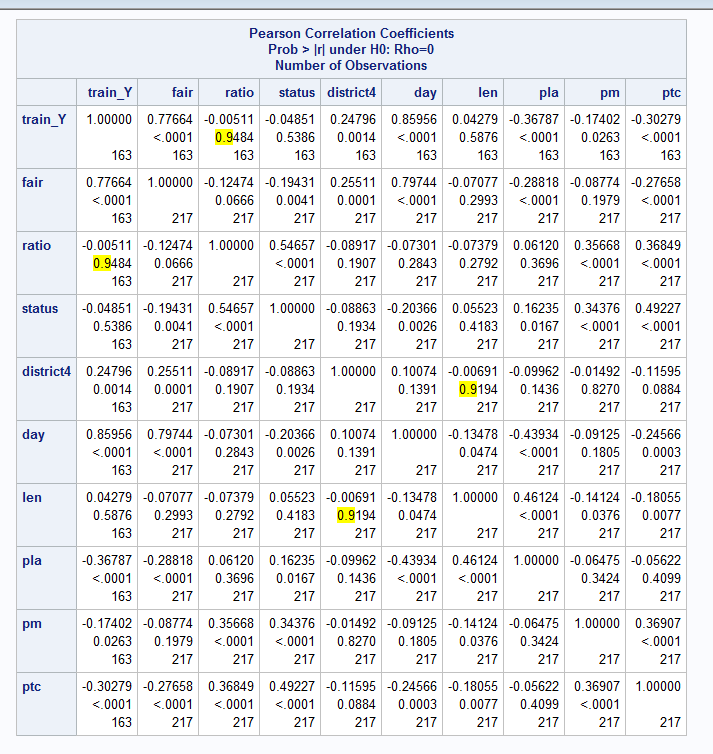
Item F3:



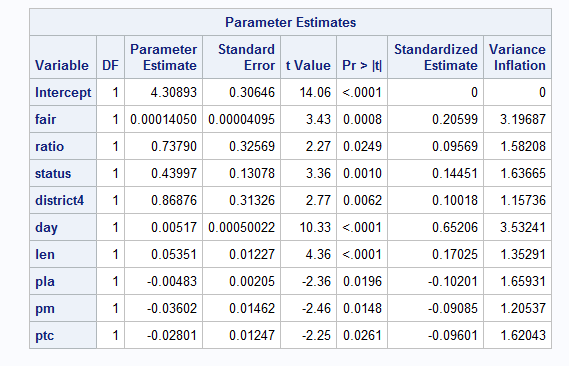
Item G1:



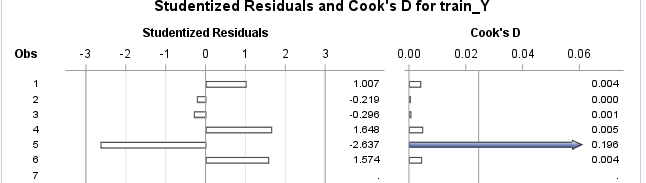
Item G2:



Item G3:



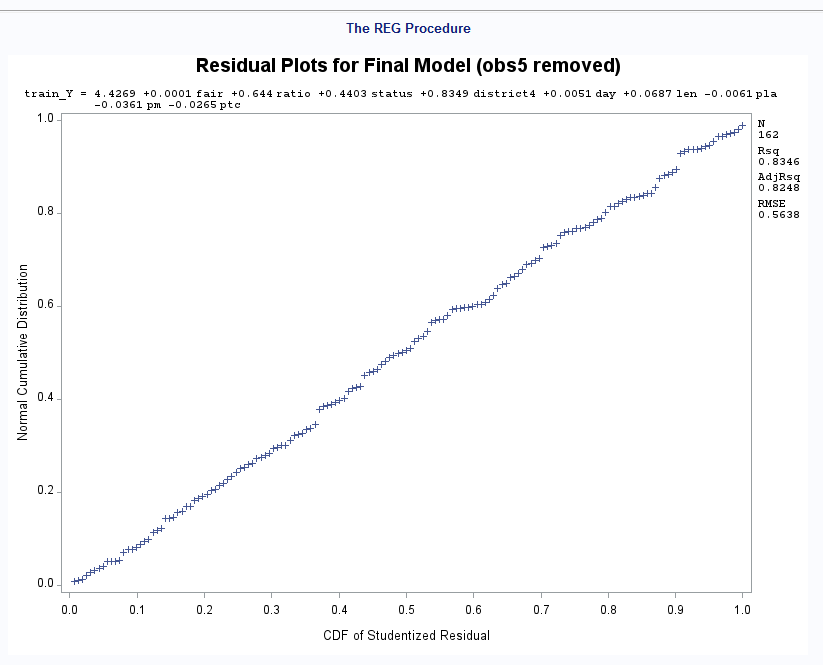
Item G4:



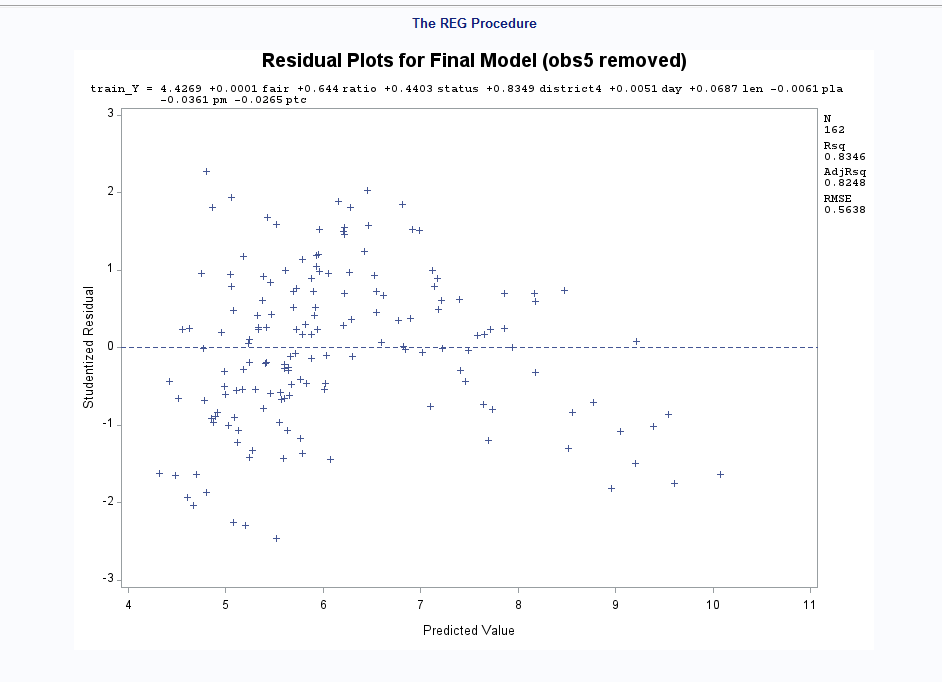
Item H1:



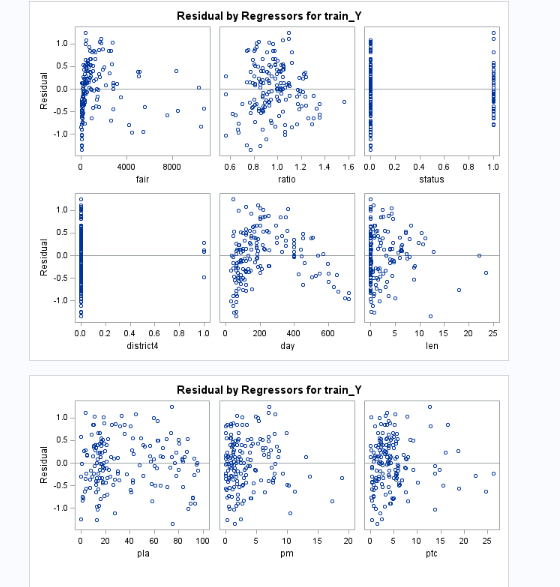
Item H2:



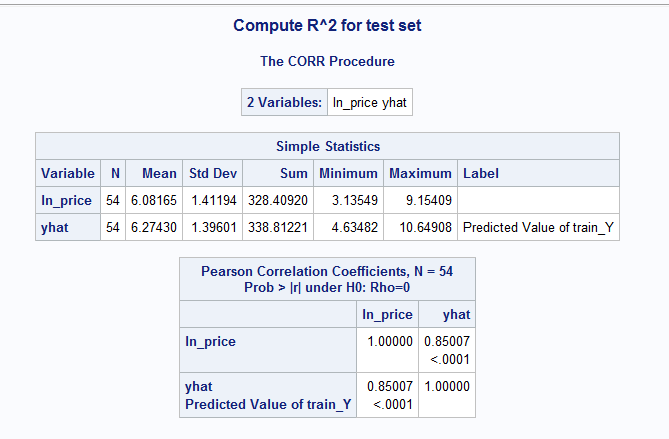
Item H3:



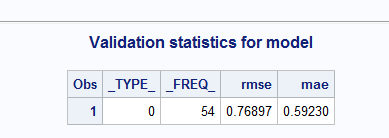
Item H4:



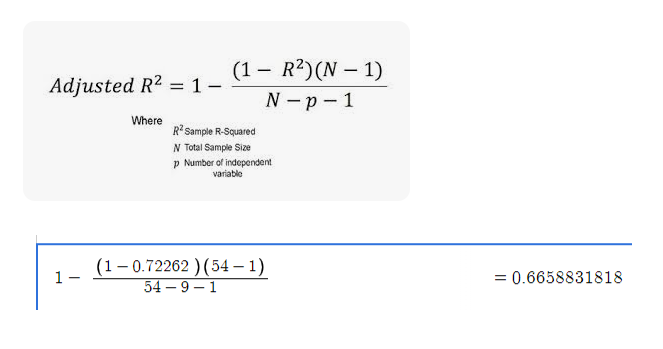
Item I1:



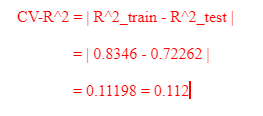
Item I2:



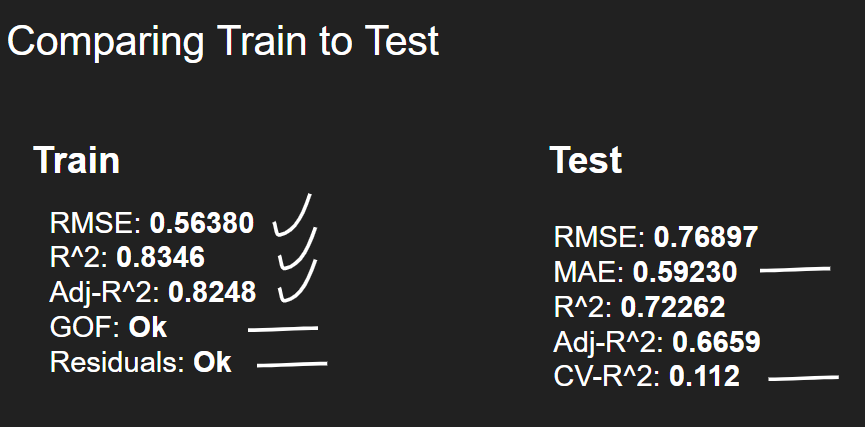
Item I3:



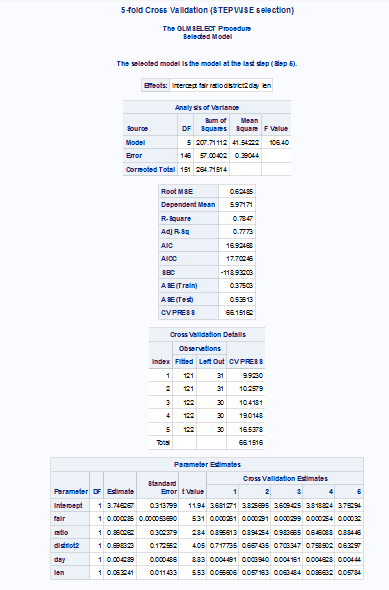
Item I4:



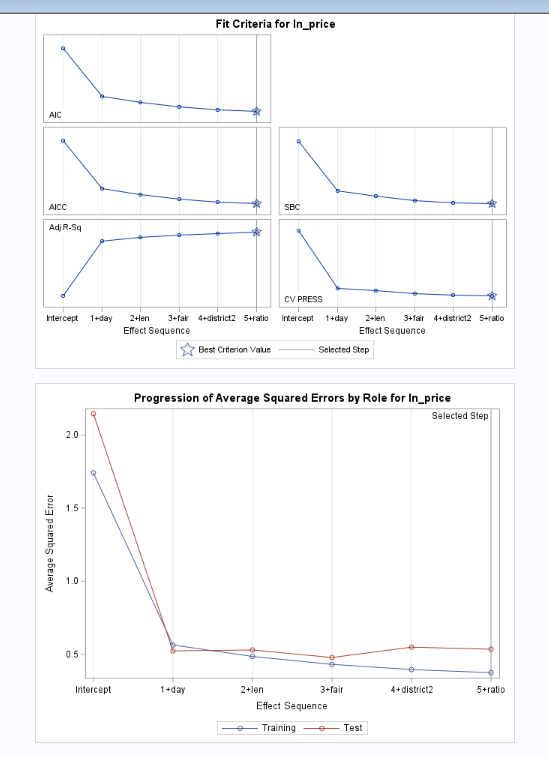
Item I5:



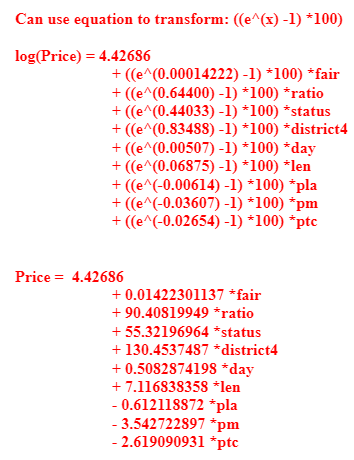
Item J1:



Item J2:



Item K:



Item L:

