INDIAN SCHOOL OF BUSINESS

Project Report

STATITSTCAL ANALYSIS & MODELLING

LINEAR REGRESSION MODELLING TO PREDICT AIR-FARES DATA

-Rajiv V

Table of Contents

Table of Figures

Data Analysis on Airplanes Data:

Null values Check:

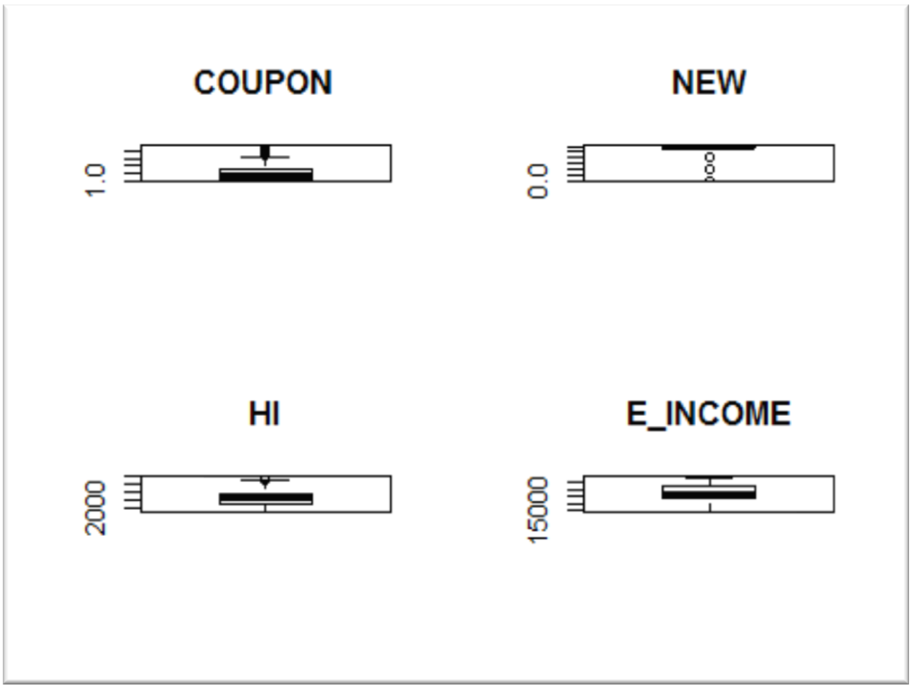
We need to check if there are any null values present in our data set and should take a decision to remove the records or to impute the Nulls with appropriate data.

We have found that there are no nulls in our dataset. So we have no problems with proceeding further.

Univar ate Analysis:

Before proceeding to the fitting of the model it is good to check how the variables are behaving independently. Box-plot will help us in analysing the univariate analysis.

Figure 1: Boxplot of Coupons, New aircarriers, Herfindel Index, Ending Cities Average Income



From Figure 1, we can see in few routes' the average numbers of coupons (COUPONS) are more than the rest of the other routes, upon inspection we can see that average number of coupons issued are more usually more in the routes of: Boston-Sandiago, Salt Lake City - Baltimore/Wash Intl, Columbus – Los Angeles etc.., The Number of new Carriers (NEW) entering few routes are low in few regions. This may be because of the less popularity of those routes. Routes such as: Oakland – Reno, Burbank - San Jose, Burbank - San Francisco are the routes which seem to have very high market concentration than other airplane routes. The income of the individuals in the cities to which the flight usually go are normally distributed.

From the below Figure 2, we can see that San Francisco is having the highest per capita and El Paso is having the lowest income from where our fight usually starts. Since SF in in California and it is usually the city with highest income in USA. The routes like: Las Vegas - Honolulu (Intl), Boston - San Francisco, Boston – San Jose are few of the longest distances through which our aircraft has to travel. The populations are normally distributed in all the cities.

Figure 2: Average income of Staring City, Distance of the route, Population of starting and ending cities

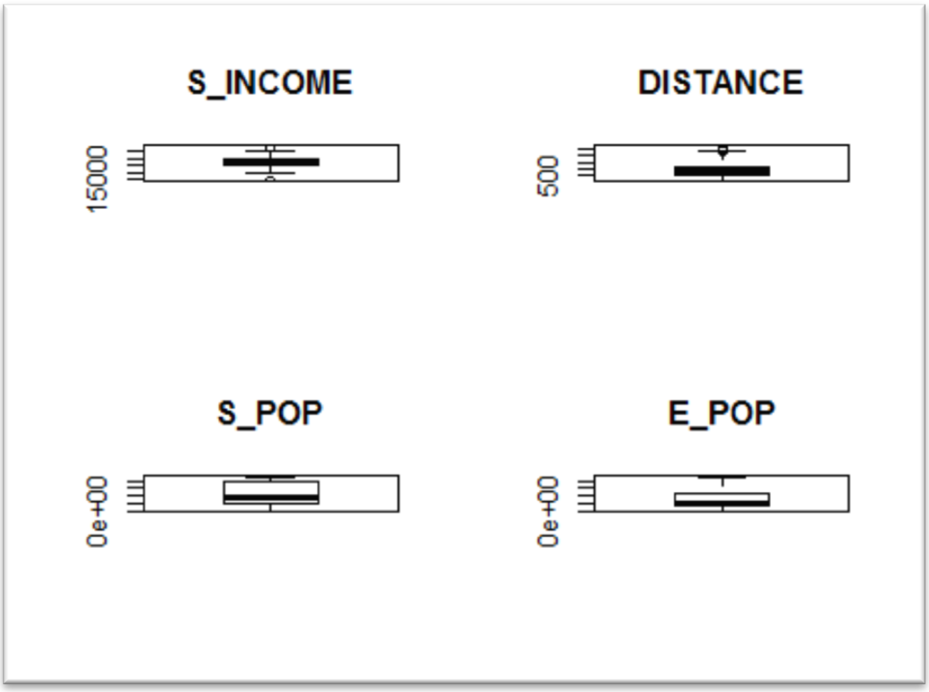
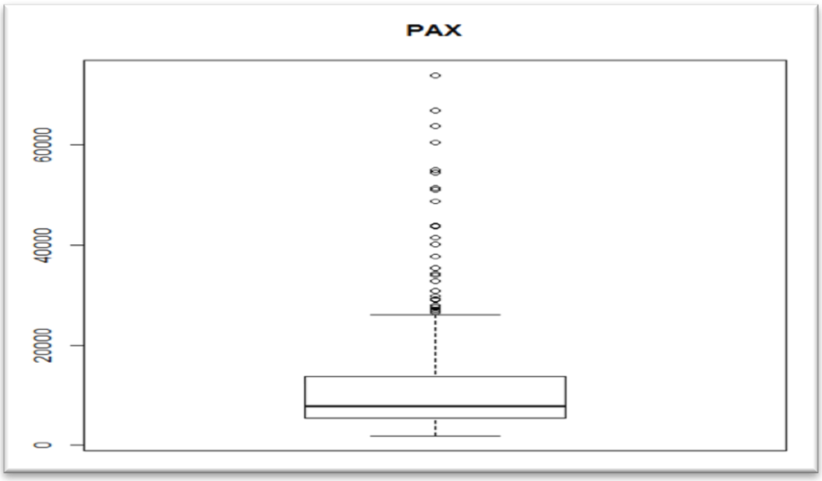


Figure 3: Number of Passengers on the route travelled.

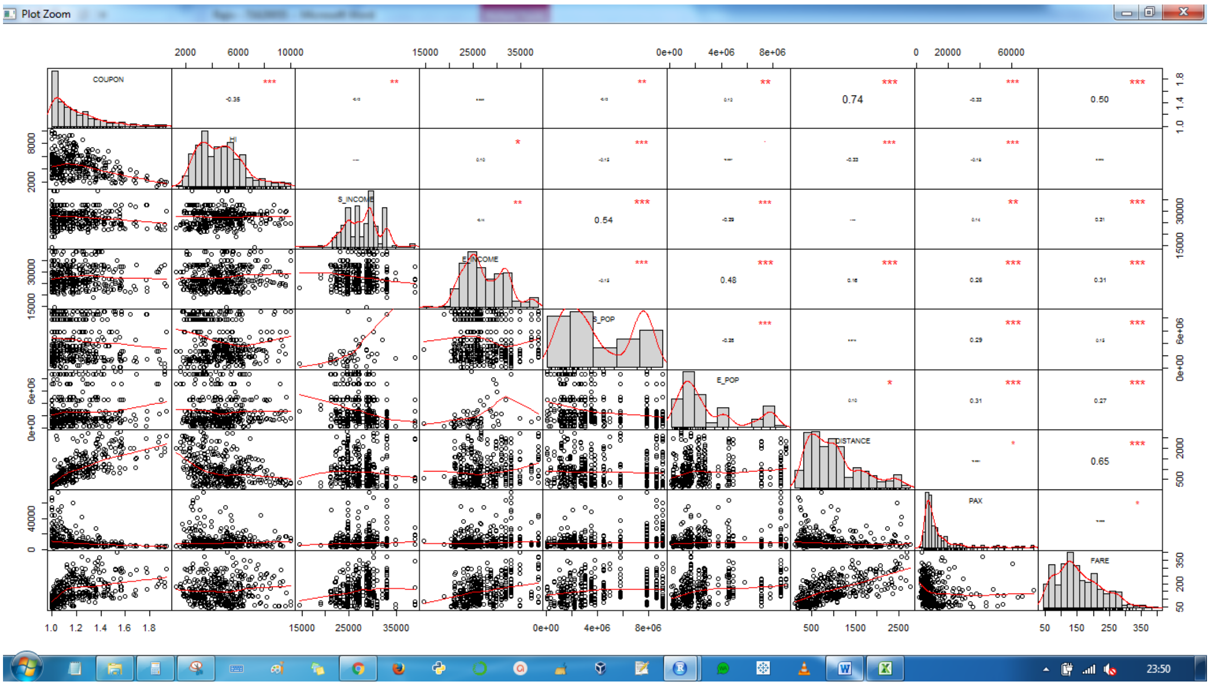


From the Figure 3, few routes which are having very high passenger traffic than the normal routes which the airlines operates. The routes are towards New York from cities like Chicago, Boston, Los Angeles. To Washington from New York etc..,

Bivariate Analysis:

We need to check how two independent regressors are correlated to each other and also how they are related to the dependant variable, the price of the ticket.

Figure 4: Scatter plot of all the variables



From Figure 4, we can see the correlation coefficients, scatter plot and the distribution of each continuous variable in our data set. The significantly correlated variables are: Fare (our dependant variable) to Distance and avg. number of coupons, coupons and distance, population and income. We can also see that none of the regressors are normally distributed in nature.

After we had initial flavour of our data, we know how our parameters are acting. So we can now go and start building the model which we will see in the next section.

Model Fitting:

We will start by assuming that fitting a linear regression model is appropriate for the predicting our air fares. We first build a basic regression model and then slowly improve our model's to get the best model that fit's for optimal solution.

Data Partitioning:

In order to evaluate the performance of different models we split the data in hand into two. One is the training data set and the other is the test/validation data set. We randomly pick 80% of the records from our and then train our model on the given data. The 20% test set is used to evaluate the best equation.

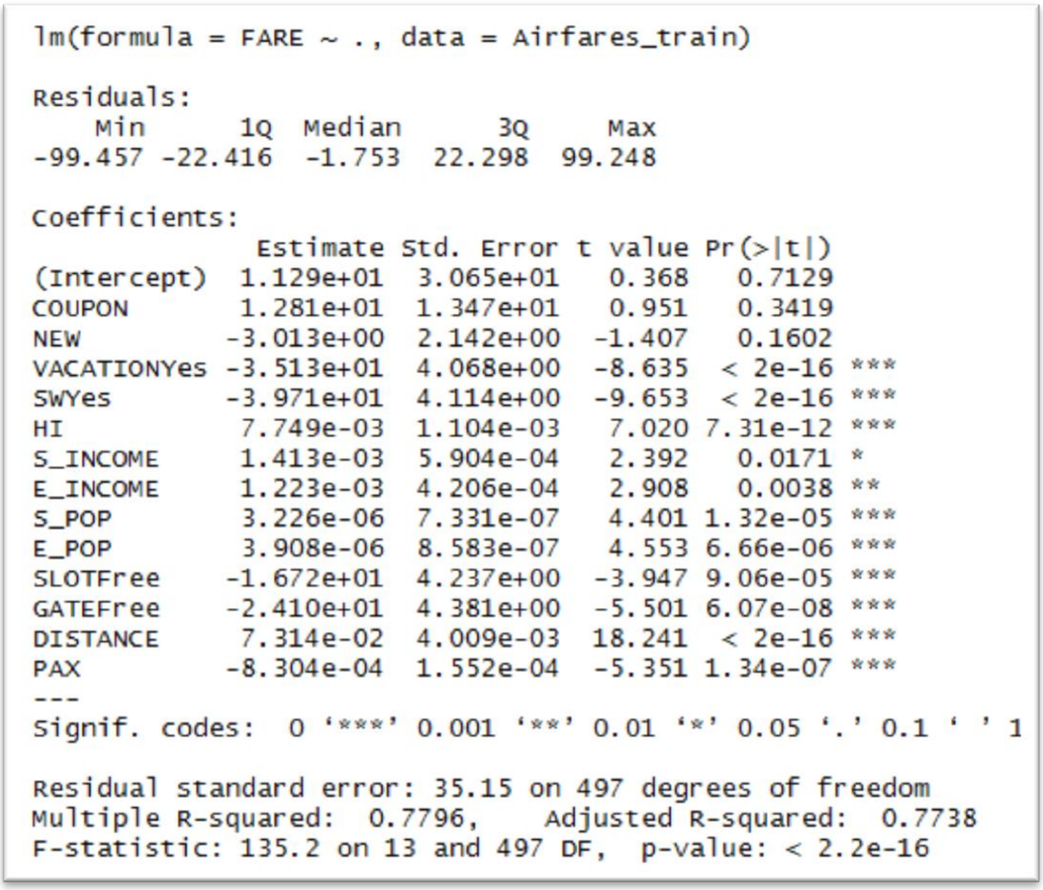
Fit 1:

This is the naïve model that we build on our (training) data. This will give us a fair idea of on how to proceed by improving our model.

fit1<-lm(FARE~.,data = Airfares_train)

Summary of Fit1 Model:

Below is the summary of the model.



```
lm(formula = FARE ~ ., data = Airfares_train)

Residuals:
    Min      1Q  Median      3Q     Max 
-99.457 -22.416  -1.753  22.298  99.248 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.129e+01  3.065e+01   0.368   0.7129    
COUPON       1.281e+01  1.347e+01   0.951   0.3419    
NEW         -3.013e+00  2.142e+00  -1.407   0.1602    
VACATIONYes -3.513e+01  4.068e+00  -8.635  < 2e-16 ***
SWYes       -3.971e+01  4.114e+00  -9.653  < 2e-16 ***
HI           7.749e-03  1.104e-03   7.020 7.31e-12 ***
S_INCOME     1.413e-03  5.904e-04   2.392   0.0171 *  
E_INCOME     1.223e-03  4.206e-04   2.908   0.0038 ** 
S_POP        3.226e-06  7.331e-07   4.401 1.32e-05 ***
E_POP        3.908e-06  8.583e-07   4.553 6.66e-06 ***
SLOTFree    -1.672e+01  4.237e+00  -3.947 9.06e-05 ***
GATEFree    -2.410e+01  4.381e+00  -5.501 6.07e-08 ***
DISTANCE     7.314e-02  4.009e-03  18.241  < 2e-16 ***
PAX         -8.304e-04  1.552e-04  -5.351 1.34e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35.15 on 497 degrees of freedom
Multiple R-squared:  0.7796,	Adjusted R-squared:  0.7738 
F-statistic: 135.2 on 13 and 497 DF,  p-value: < 2.2e-16
```

Figure 5: Summary of the Fit1 Model

On the Top line in the Figure 5, we can see the five point distribution of our residuals (Actual value Minus Predicted value) that we got by fitting this model.

Next we can see the different estimators (Beta values) for all our regression parameters.

The intercept is usually not taken into consideration and it's interpretation is usually not mostly useful.

The estimates of the regressors will say how each of the regressor is contributing/ influencing in determining our airfare. The positive sign of the estimate says that the variable is contributing in a positive way towards our output variable and vice-versa. From inspection we can say, income, population, distance, avg. coupons are more the airfares to-

fro will be usually more. If the route is an vacation route, South west airlines operating in the route, number of passengers travelling will reduce our fares with the order of magnitude of estimate.
The std. Error says the range in which our estimates will fall above and below the point estimate we got.
The t-value and P-values will say if each regressor is actually significant at a particular confidence level in determining our predictor variable. We cans see different *'s mapped across each regressor which says the significance of that regressor at different confidence intervals which are given below. The estimator variables like avg. coupons, The number of new carriers entering that route are not significant in determining our airfare and the rest are significant at various levels of significance.

"Residual standard error", will say the interval in which our each output variable will fall. This should be usually as small as possible because the smaller the standard error the smaller is our prediction uncertainty in our output variable. Here our standard error value is 35.15. which is saying that our predicted airfare will fall in the interval above and below the point estimate which we get.

We can also see the "Multiple R-squared" value which came out as 0.7796 is saying that almost 78% of the total variation in the given data has been captured by the basic model which we have built now. "Adjusted R-squared", will say us how much our model accuracy can be improved if we can add one more regressor. It is usually close to and less than the R-squared value.

The F-statistic and the corresponding p-value, are the values derived from the Hypothesis test (ANOVA test) of fitting a linear model to this set of data is appropriate or not. We can see that our p-value is very small even at 1%. So, we say that we reject our null hypothesis at this level of significance and proceed with our alternate hypothesis that a linear model is appropriate for our data in hand.

We have now seen how do we read our regression summary output. We will now test our assumptions that we have before building our model are holing true after building our model or not.

QQ-Plot of residuals:

One of our major assumption in linear modelling is that all our residuals are normally distributed. The predictions are all random and are not related to each other.
The best plot to determine a normality assumption of a variable is to check it's qq-plot. If all our points or most of our points will fall within our confidence bands then we can say our variable is normally distributed.

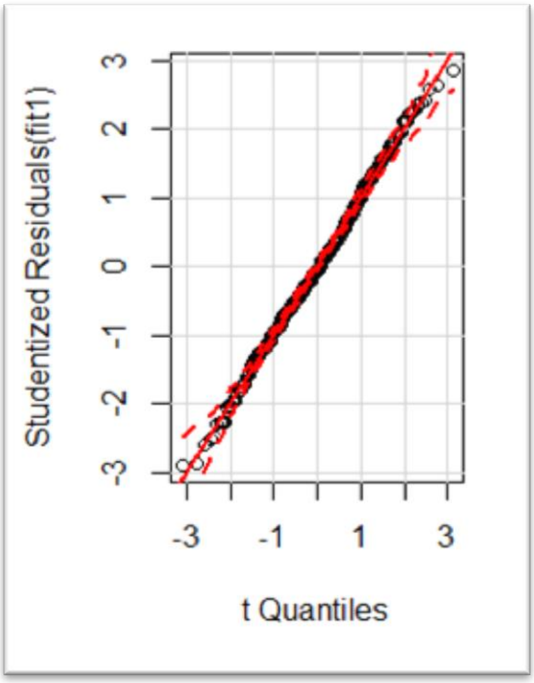


Figure 6: QQ-Plot of Fit1 model.

From the plot, Figure 6, for residual generated will say that there are few points which are close to the bands but are falling outside the confidence bands.

Residuals Vs Regressors plots:

The other important assumption is that, we are that the regression variables will not have any influence on the estimated values or the residual values. So, when we plot the residual vs regressor plot for our model we should not see any kind of trend or pattern from the plotted graph.

The below figure 7 & Figure 8 says that, the relation between the regressors and the residuals.

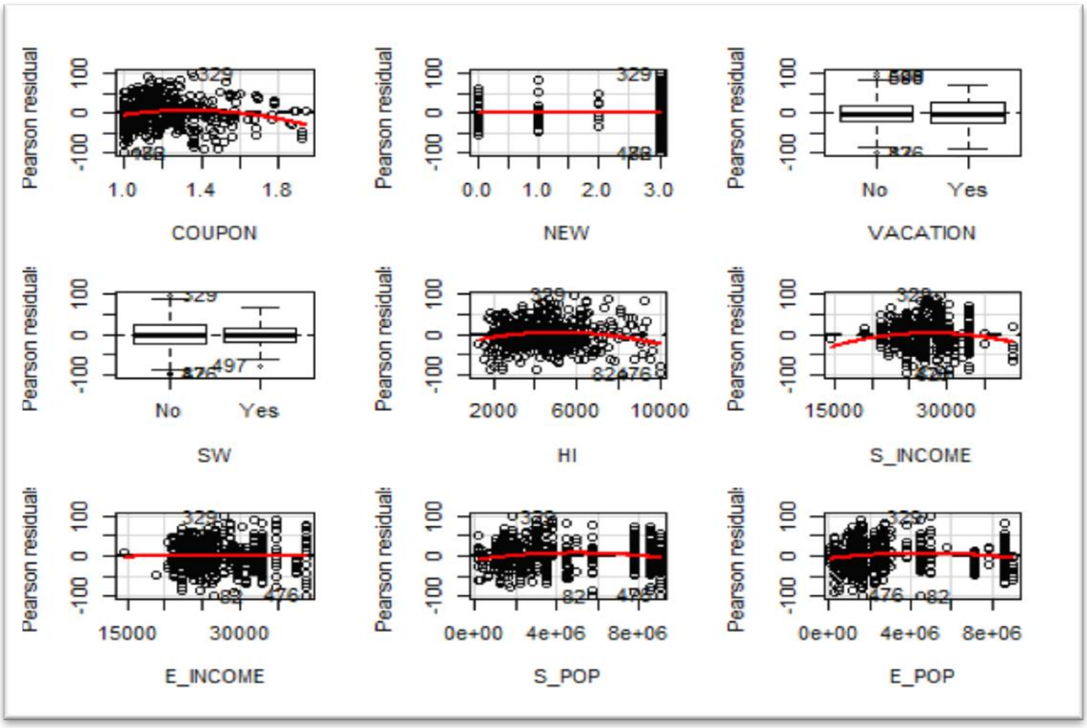


Figure 7: Regressors vs residuals plot1 for Fit1 model

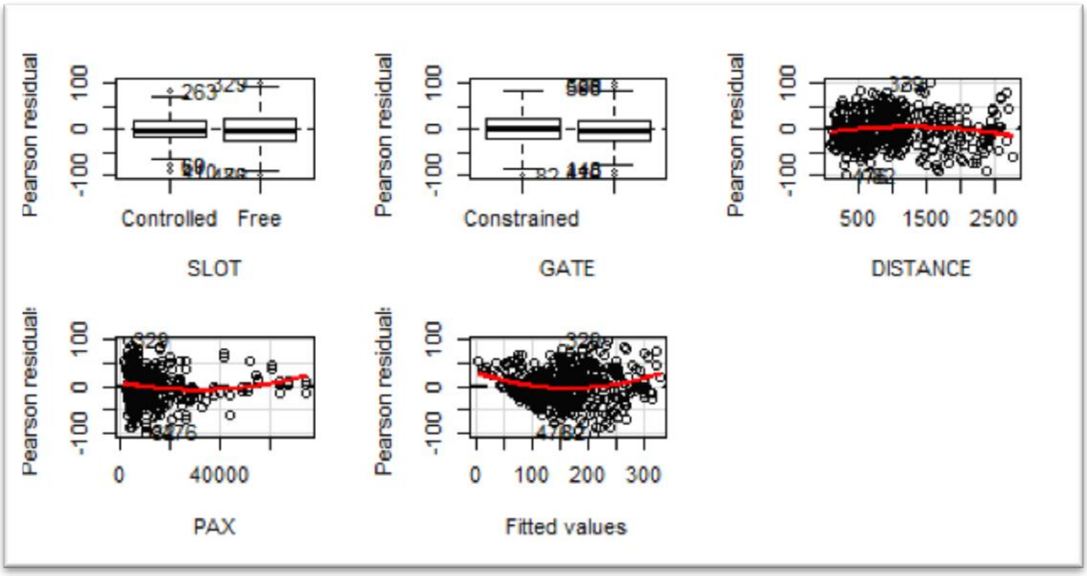


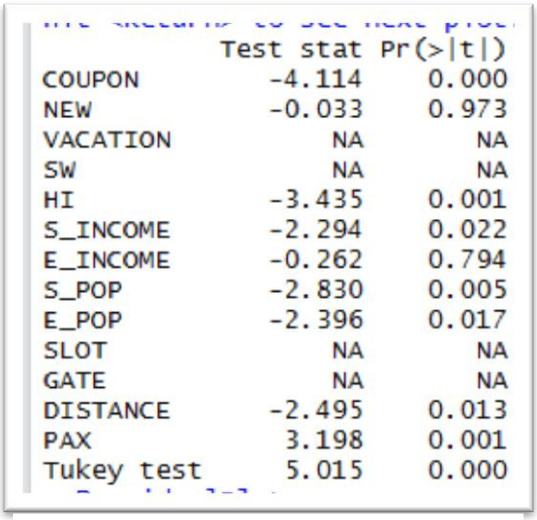
Figure 8: Regressors vs residuals plot2 for Fit1 model

The red line in each plot will be able to say us if that parameter is having any kind of influence on the predicting our output variable. If the line is straight, we can say that that there is no influence of that parameter on the residuals. We need to examine each plot to determine if there is any curvature in the line which will say the kind of the influence it has.

Upon examination we can see: Avg. coupons (COUPONS), measure of market concentration (HI), avg. income of in the starting city, population in the starting and ending city, distance and number of passengers are all having a quadratic influence on our output variable.

Curvature Test:

This curvature test is the same output of the above plots we described for residuals vs regressors. This test will check the hypothesis if each regressor is having a quadratic influence on our regression output. The below table in the below figure 9 will give the p-values of the t-statistic. If the P-value is less than 5% then we can say that the regressor has a quadratic influence on our output. We can see the same variables specified above in regressor's vs residuals plot are having less p-value.



```
              Test stat Pr(>|t|)
COUPON           -4.114    0.000
NEW              -0.033    0.973
VACATION             NA       NA
SW                   NA       NA
HI               -3.435    0.001
S_INCOME         -2.294    0.022
E_INCOME         -0.262    0.794
S_POP            -2.830    0.005
E_POP            -2.396    0.017
SLOT                 NA       NA
GATE                 NA       NA
DISTANCE         -2.495    0.013
PAX               3.198    0.001
Tukey test        5.015    0.000
```

Figure 9: Curvature Test for Fit1 model

Fit2:

As we have seen that there are some drawbacks in the above model because they did not hold to our initial assumptions of building the model we will now try to address few of the concerns and improve the prediction accuracy of the output.

Transformation of variables:

← Do →

Summary of Fit2 Model:

← Do →

QQ-Plot of residuals:

← Do →

Residuals Vs Regressors plots:

← Do →

Check of Influential Observations:

← Do →

Model Validation:

← Do →

Final Model:

← Do →

Summary of regression process:

Conclusion / Business Insights from regression:

← Do →