Deregulation of the Intrastate Trucking Industry

Kevin Merritt, Valeria Torres, Frances Chua

California State University, Long Beach

STAT 410/510

Dr. Tianni Zhou

ABSTRACT

In the early 1980's, several states removed regulatory restraints on shipping prices. Research was done on Florida's transportation with two goals in mind: Assess the impact of deregulation on prices charged for motor transport service and creating a model for predicting future prices. Variables for the model were created from factors which were thought to be significant in transportation pricing. A subset of variables were predicted as significant using the adjusted R^2 and Mallow's C_p Criteria and then validated using a Stepwise Regression model selection. There was a single observation that was an influential outlier but its removal from the analysis led to the same conclusions. Deregulation was a significant factor in reducing shipment prices, and led to adding it to a model predicting future prices.

INTRODUCTION

Data evaluated were gathered from 27,000 individual shipments from the largest carriers in Florida taken before and after deregulation. The particular cases studied are from a particular carrier whose trucks originated in either Jacksonville or Miami. The dependent variable is the price charged per ton-mile denoted by y. The independent variables used to predict y are as follows: x_1=distance, x_2=Weight, x_3=Pctload, x_4=Origin, x_5=Market, x_6=Dereg, x_7=Carrier, and x_8=Product. The motivation of analyzing this data is to assess the impact of deregulation on price charged for transport services in Florida, and to estimate a model of the supply price for predicting future prices.

1. STANDARD PROCEDURES

In any statistical analysis there are two things that must be checked called "simple statistics": Normality, and correlation among variables. Normality of the response variable will be checked by observing a Q-Q plot, while multicollinearity would be seen by checking a correlation matrix. Before applying these procedures there were items needed to be addressed. First the dependent variable x_7=Carrier was immediately deleted from the analysis since there was no changes for any of the observations (all input for observations was B). Lastly there was categorical transformation for the following: x_{4n}=Origin where "MIA"=1 and "JAX"=0, x_{5n}=Market where "LARGE"=1 and "SMALL"=0, x_{6n}=Dereg where "YES"=1 and "NO"=0.

1.1 NORMALITY ASSUMPTION

A Q-Q plot is a graphical tool used to check if it is plausible that our data came from a normal distribution. The desired outcome is a fairly straight line. In figure 1, it can be clearly seen that plotting $y1$=PRICPTM against Normal quantiles yields an exponential graph. By applying a log transformation, seen in figure 2, the data has now been normalized. Therefore from now on the log transformation of prices will be used going forward denoted$y2$.

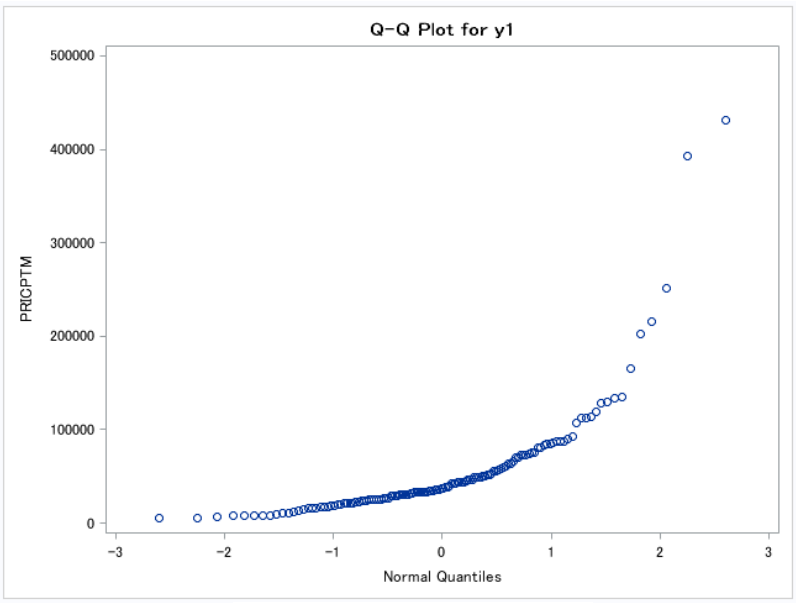

Figure 1: Q-Q plot for PRICPTM (y1)

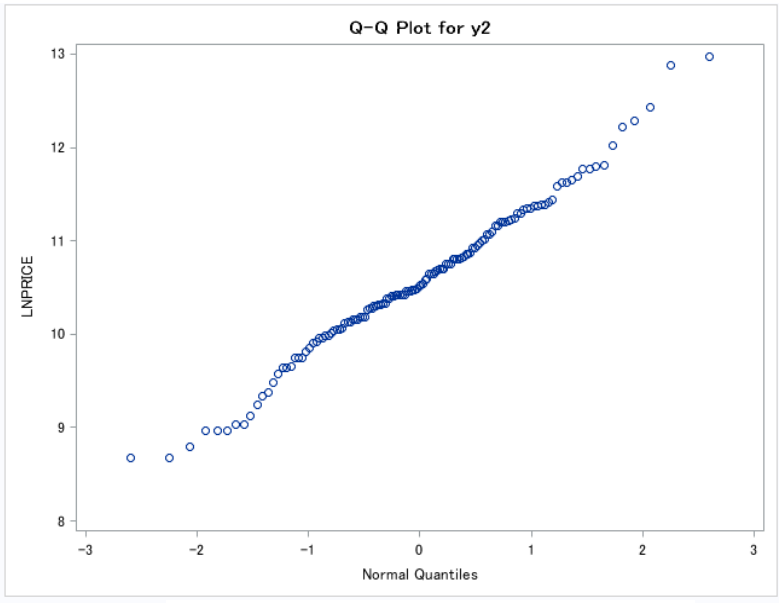

Figure 2: Q-Q plot for log transformation of PRICE

1.2 CORRELATION BETWEEN VARIABLES

The correlation coefficient (Table 1) shows us that x_2=Weight and x_3=Pctload are extremely correlated, the value being .99967. To avoid problems of multicollinearity the variable x_2 will be removed while x_3 stays in the model for the rest of the analysis.

Pearson Correlation Coefficients, N = 134							
	x1	x2	x3	x4n	x5n	x6n	x8
x1 DISTANCE	1.00000	0.06964	0.06798	0.08887	-0.21937	-0.05898	0.05352
x2 WEIGHT	0.06964	1.00000	0.99967	0.01301	0.04699	-0.05374	-0.02928
x3 PCTLOAD	0.06798	0.99967	1.00000	0.01484	0.04656	-0.05414	-0.02774
x4n ORIGINNUM	0.08887	0.01301	0.01484	1.00000	0.08243	0.10466	0.03152
x5n MARKETNUM	-0.21937	0.04699	0.04656	0.08243	1.00000	0.03012	-0.04575
x6n DEREGNUM	-0.05898	-0.05374	-0.05414	0.10466	0.03012	1.00000	0.07146
x8 PRODUCT	0.05352	-0.02928	-0.02774	0.03152	-0.04575	0.07146	1.00000

Table 1: Pearson Correlation Coefficients

2. METHODS

The goal of model selection is to choose a simple model that adequately explains the data. The final estimated model would be able to predict future prices, the response variable, in relation to our predictor variables, the chosen $x_k \; for \; k = 1, \dots, 8$. There are three methods being considered for estimating a model that predicts future prices: *Adjusted R-Square, Mallows C_p Criteria,* and *Stepwise regression.*

2.1 METHOD SELECTION

The first two methods are $R_{a,p}^2$ *(adjusted R-square)* and *Mallows C_p Criteria* which are used in tandem. $R_{a,p}^2$ measures the proportion of variation explained by only those independent variables(x_k) that really affect the dependent variable. It penalizes for adding independent variables that do not affect the dependent variable. The point in which the addition of more of these x_k will cause $R_{a,p}^2$ to level off is the desired point. *Mallows C_p* is calculated for all possible subset models. Using this technique, the model with the smallest $C_p = \frac{SSE_p}{MSE(x_1,\dots,x_{p-1})} - (n - 2p)$ is declared the best linear model. As the number of independent variables x_k increases, an increased penalty term ($2p$) has a decreased SSE. The desired value will be where $C_p \leq p$ which will give the desired subset of predictors. In short $R_{a,p}^2$ should be maximized while C_p should be minimized.

The third method *Stepwise Regression* is used to further validate the findings of $R_{a,p}^2$ and *Mallows C_p Criteria*. This method begins with a null model then enters and removes predictors, in a stepwise manner, until there is no justifiable reason to enter or remove any other predictors.

Lastly, *Graphs* of the residuals against each predictor will be used to see if higher order terms will improve the model. Two-way *interactions* within the different variables will be tested to see if there can be an addition of an interactive term in our model. Significance will be conducted by *hypothesis testing.*

2.2 IDENTIFYING EXTREME VALUES

To find outlying observations, the *Studentized Residuals* of each observation will be assessed to check for any outlying observation in the y -direction. Values above

$t\left(1 - \frac{\propto}{2}; n - p - 1\right) = t\left(1 - \frac{.05}{2}; 134 - 7\right) = 3.656$ will be considered outliers. The *leverages* (h_{ii} of the hat matrix) for each observation will be checked to find any outliers in the x-direction. Leverages are considered large if they are greater than twice the mean leverage value ($\frac{2p}{n} = \frac{2*7}{134} = .10448$).

For influential cases *DFFITS, Cook's Distance*, and *DFBETAS* of each outlying observation will be used to check their influence. *DFFITS* looks at the difference in fitted values for a single observation. This if found by looking at values greater than $2\sqrt{\frac{p}{n}} = 2\sqrt{\frac{7}{134}} = .45712$. *Cook's Distance* looks at the influence of fitted values for all observations with its associated $F_{p,n-p}$ distribution statistic for influenital variables. If the values found by *Cook's Distance* are greater than 50% then they will be considered influential. *DFBETAS* looks at the influence on regression coefficients. These values are considered highly influential is greater than $\frac{2}{\sqrt{n}} = \frac{2}{\sqrt{134}} = 0.17277$.

3. RESULTS

3.1 MODEL SELECTION

3.1.1 SELECTION STATISTICS

Selection statistics can be used to determine which model is the best fit for the data. Below (table 2) we have outputted the best model for each number of predictors. It is easy to see that the model with five predictors has the highest adjusted r-square value and the lowest AIC and BIC values. For Mallow's Criterion, the model with six predictors has the value closest to the number of predictors in the model but the model with five predictors is the lowest value and is still rather close to the number of predictors. Since all the other criterion point to the model with five predictors, we would choose to use the model with x_1, x_2, x_{4n}, x_{6n} and x_8.

Number in Model	R-Square	Adjusted R-Square	C(p)	AIC	BIC	SSE	Variables in Model
1	0.2969	0.2916	468.0690	-96.5013	-98.8093	63.29649	x1
2	0.6096	0.6036	204.1017	-173.3288	-175.6195	35.14790	x1 x6n
3	0.7993	0.7946	44.7519	-260.4696	-260.2842	18.07148	x1 x2 x6n
4	0.8355	0.8304	15.9388	-285.1365	-283.5645	14.81039	x1 x2 x4n x6n
5	0.8519	0.8461	4.0008	-297.1954	-294.4416	13.33526	x1 x2 x4n x6n x8
6	0.8519	0.8449	6.0003	-295.1959	-292.3151	13.33521	x1 x2 x4n x5n x6n x8
7	0.8519	0.8436	8.0000	-293.1962	-290.1884	13.33518	x1 x2 x3 x4n x5n x6n x8

Table 2: Selection Statistics

3.1.2 STEPWISE SELECTION

Another method for choosing the correct model is to use stepwise regression. For our stepwise selection, we chose to set the entering significance level to $\alpha = .10$ and the significance level to leave at $\alpha = .15$. The summary of the stepwise selection process for our data is below (table 3). We can see from the table below that x_1 is the first predictor added to the model and x_8 is the last one added. The rest are shown in the order that they were added to the model. Looking through the steps of the stepwise also reveals that at no point was a

predictor dropped from the model. This is confirmed to be the correct choice by the fact that all of the predictors p-values are well below the exiting significance level of $\alpha = .15$. Therefore, it is clear that the model chosen from the selection statistics and from the stepwise process are the same.

Summary of Stepwise Selection									
Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x1		DISTANCE	1	0.2969	0.2969	468.069	55.74	<.0001
2	x6n		DEREGNUM	2	0.3127	0.6096	204.102	104.91	<.0001
3	x2		WEIGHT	3	0.1897	0.7993	44.7519	122.84	<.0001
4	x4n		ORIGINNUM	4	0.0362	0.8355	15.9388	28.40	<.0001
5	x8		PRODUCT	5	0.0164	0.8519	4.0008	14.16	0.0003

Table 3: Stepwise Selection Summary

3.1.3 CHECKING ASSUMPTION OF EQUAL VARIANCES

Now that the model has been chosen, we need to check these variables for equal variances to verify that the assumption is not violated. Below (figure 3), we can see that four of the five predictors chosen for the model do not violate the assumption of equal variances. However, the variable distance (x1) has a very distinct parabolic shape. This indicates an ${x_1}^2$ (x1sq) term should be added to the model. After adding x1sq we can see that the residuals vs distance plot (figure 4) has evened out and no assumptions of equal variances are violated anymore.

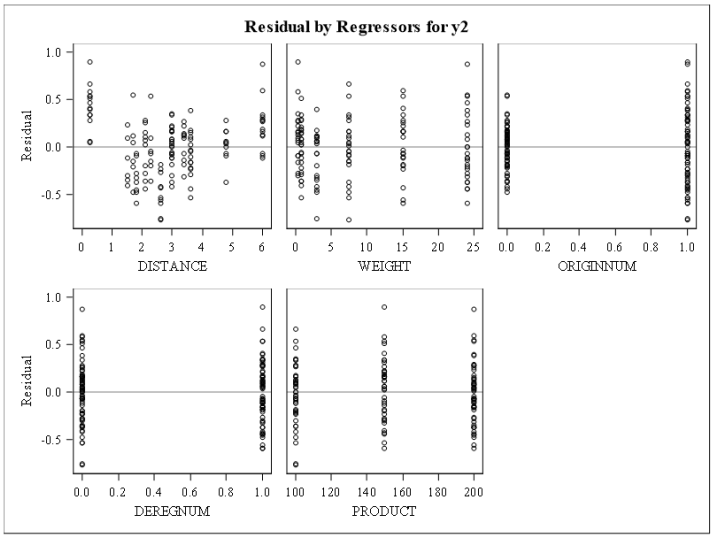

Figure 1: Residuals vs Predictors Before Adding X^2

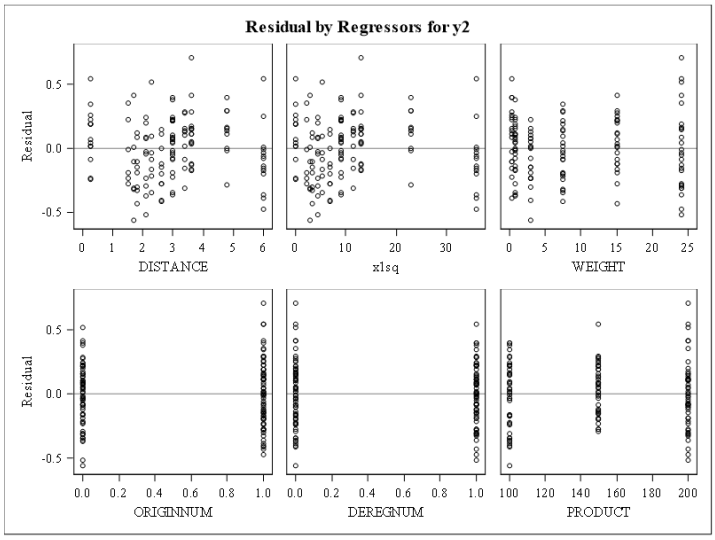

Figure 4: Residuals vs Predictors After Adding X^2

Now that we have added a new predictor to our model, we need to verify that all predictors in the model are still significant. Looking at the table below (table 4), we can see that variable x_{4n} now has a p-value of .8275. This is well over the alpha level of .15 and hence can be removed from the model because it is no longer significant. The new set of predictors used in the model are $x_1, {x_1}^2, x_2, x_{6n}$ and x_8.

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	12.44913	0.11769	105.78	<.0001
x1	DISTANCE	1	-0.80962	0.05498	-14.73	<.0001
x1sq		1	0.08075	0.00851	9.49	<.0001
x2	WEIGHT	1	-0.04127	0.00244	-16.91	<.0001
x4n	ORIGINNUM	1	0.01171	0.05364	0.22	0.8275
x6n	DEREGNUM	1	-0.97768	0.04341	-22.52	<.0001
x8	PRODUCT	1	0.00308	0.00051842	5.95	<.0001

Table 4: Significance Test with X^2

3.1.4 TWO-WAY INTERACTION TERMS

The next step is to check if any interaction terms should be added to the model. There are six possible interaction terms; $x_{12}, x_{16n}, x_{18}, x_{26n}, x_{28}, x_{6n8}$. Running the model with each of these predictors added in separately we will find that the only two that are significant are x_{26n} (table 5) and x_{6n8} (table 6). However, looking at the model with x_{6n8} added, we see that x_{6n} becomes a nonsignificant predictor in the model. We can't add in an interaction term that doesn't include the parent predictor variables so x_{6n8} will not be added to the model. Therefore, we only add in the two-way interaction term x_{26n}.

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	12.32493	0.08832	139.55	<.0001
x1	DISTANCE	1	-0.84720	0.03804	-22.27	<.0001
x1sq		1	0.08676	0.00570	15.22	<.0001
x2	WEIGHT	1	-0.02657	0.00276	-9.61	<.0001
x6n	DEREGNUM	1	-0.69396	0.05088	-13.64	<.0001
x8	PRODUCT	1	0.00331	0.00042669	7.75	<.0001
x26n		1	-0.03136	0.00405	-7.75	<.0001

Table 5: Significance Test with x26n Added

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	11.98090	0.11035	108.58	<.0001
x1	DISTANCE	1	-0.81103	0.03861	-21.01	<.0001
x1sq		1	0.08188	0.00578	14.16	<.0001
x2	WEIGHT	1	-0.04050	0.00205	-19.71	<.0001
x6n	DEREGNUM	1	-0.01072	0.13795	-0.08	0.9382
x8	PRODUCT	1	0.00613	0.00060374	10.16	<.0001
x6n8		1	-0.00630	0.00086792	-7.26	<.0001

Table 6: Significance Test with x6n8 Added

3.1.5 FINAL MODEL

The final model is:

$$Y = \beta_1 X_1 + \beta_1{}^* X_1{}^2 + \beta_2 X_2 + \beta_6 X_{6n} + \beta_{26} X_{26n} + \beta_8 X_8$$

The final predicted equation is:

$$Price = 12.325 - .847 * distance + .087 * distance^2 - .027 * weight$$
$$-.694 * deregnum + .031 * weight * deregnum + .003 * product$$

3.2 OUTLIERS

3.2.1 OUTLYING OBSERVATIONS

In our case, an observation is considered outlying based on studentized residuals if it is larger than 3.656. Looking at the values in the complete output (not included) we see that there are no values for Rstudent that are larger than 3.656. The leverages (HatDiagonal) of each observation can be reviewed to check for outliers in the x direction. Leverages are considered large if they are larger than twice the average leverage of .10448. The table below (table 7) shows the observations that have a large HatDiagonal value and need to be checked to see if they are influential.

Obs	RStudent	CooksD	HatDiagonal	DFFITS	DFB_Intercept	DFB_x1	DFB_x1sq	DFB_x2	DFB_x6n	DFB_x8	DFB_x26n
4	1.4573	0.041	0.1195	0.5369	0.3065	-0.2739	0.2136	0.0112	-0.0696	-0.1550	0.2347
9	0.3577	0.003	0.1227	0.1338	0.0310	-0.0235	0.0457	-0.0034	-0.0121	-0.0460	0.0580
53	-1.3686	0.035	0.1160	-0.4958	0.0879	0.1204	-0.2009	-0.0064	0.0478	-0.1286	-0.2014
72	-0.2023	0.001	0.1166	-0.0735	-0.0367	0.0458	-0.0372	-0.0441	-0.0101	0.0146	0.0314
78	-1.2528	0.028	0.1126	-0.4462	-0.0723	0.1304	-0.2062	-0.2511	-0.0796	0.1177	0.1930
117	1.5126	0.044	0.1192	0.5566	-0.1404	-0.1921	0.2793	0.3259	0.0947	0.1941	-0.2558

Table 7: Outlier Statistics for the Final Model

3.2.2 INFLUENTIAL OBSERVATIONS

Looking at DFFITS, we see that both observation 4 and 117 have DFFITS values above .45712. Looking at the DFBETAS for these two observations, we see that observation 4 and 117 both have DFBETA values that are greater than .17277. Checking Cook's Distance, we find that there are no observations near the 50th percentile (table 8). Therefore, we would conclude there are no influential outliers based on Cook's Distance. However, it is interesting to note that the two with the highest percentile according to Cook's Distance are observations 4 and 117. Overall, we would conclude that both observation 4 and observation 117 are influential outliers in our data set.

Observation	Cooksd	F-Value	Percentile
4	.041	.007558	.159
9	.003	.000013	.023
53	.035	.005203	.141
72	.001	.000001	.010
78	.028	.003056	.119
117	.044	.008918	.168

Table 8: Cook's Distance Calculations for Outlying Observations

3.3 CONCLUSION

In order to assess the impact of deregulation of the prices charged for motor transport services in Florida, a regression model was constructed to predict future prices. To create a more accurate model, the following factors were used in calculation: distance, weight, load capacity, city of origin (Miami or Jacksonville), and size of market destination (large or small), whether or not price deregulation was in effect, and the value of the products being shipped. When the automatic Stepwise regression method, a model was constructed that agreed with both adjusted R^2 and Mallow's C_p Criteria. The independent variables that were most significant were distance, which had the greatest effect on shipping prices, and the higher level term for distance improved estimation. The interaction term between deregulation and load capacity were also significant. The analysis determined that deregulation had an overall negative effects on shipping prices, which can be seen by the negative weight in contributes to the predicted price model. A single observation was identified as influential but with its removal the analysis led to the same results. The final model explains 93.82% of variability in the sample, so it has great prognostic value.