ANALYSIS OF THE EFFECTS OF UNIT TRAIN PRODUCTION ON RAILROAD COSTS

By

Denver D. Tolliver
Research Associate

UGPTI Staff Paper No. 58
December 1983

ANALYSIS OF THE EFFECTS OF UNIT TRAIN PRODUCTION ON RAILROAD COSTS

BY

DENVER D. TOLLIVER
RESEARCH ASSOCIATE

UPPER GREAT PLAINS TRANSPORTATION INSTITUTE
NORTH DAKOTA STATE UNIVERSITY
P. O. BOX 5074
FARGO, NORTH DAKOTA 58105

DECEMBER 1983

I. PROBLEM STATEMENT

The purpose of this analysis was to attempt to isolate the effects of unit train output on railroad costs.

Although unit train production has become increasingly important to the railroad industry in recent years, particularly with the increased demand for low-sulphur western coal during the last decade, no statistical analysis of the effects which unit train production may have on rail costs has been conducted to-date. The application of statistical techniques to railroad cost and production data, therefore, with the specific intent of measuring or capturing unit train effects may produce useful results in the areas of regulatory policy, railroad pricing, and/or long-run logistical planning by rail shippers.

II. DATA BASE

The data base used in this analysis consisted of railroad operating cost and production data for Class I American railroads. This constitutes a verified data file reported to the Interstate Commerce Commission annually by all Class I carriers.

Table 1 depicts some of the major cost and production measures which were available for this study. As Table 1 indicates, two levels of operating costs were available: (1) total operating expenses which include all aspects of cost except return on investment, and (2) various functional measures of cost, such as car costs and yard expenses, all of which add up to total operating costs. Production measures consisted of a range of output variables which were either: (1) distance measures, (2) time measures, (3) weight measures, or

TABLE 1. MAJOR RAILROAD COST AND PRODUCTION MEASURES AVAILABLE FROM ICC DATA FILES.

Cost Measures	Production Measures
Total Operating Expenses	Car Miles
Maintenance of Way Expenditures	Locomotive Unit Miles
Car Repair, Maintenance and Ownership	Train Miles
Locomotive Repair, Maintenance and Ownership	Hours Yard Switching
Train Operating Costs (Wage & Non-Wage)	Road Train Hours
Yard Operating Costs (Wage & Non-Wage)	Gross Ton Miles of Cars and Contents
Transportation Expenses	
General Administration/Overhead	Tons of Freight Originated

(4) a combination of weight and distance. Of the output measures shown, car miles, locomotive miles and train miles may be sub-divided into unit train as opposed to non-unit train output on the basis of the statistics provided.

III. MODEL FORMULATION

In approaching the problem, a two-step procedure was followed. First, the highest possible explanatory cost model was devised based on the output measures shown in Table 1. Having developed this model, the second step in the analysis was to introduce a unit train variable into the equation, and seeing the effect which this might have on the cost model and the other explanatory variables.

Independent Variable Identification

Of the output measures shown in Table 1, each might be thought to exert a substantial influence over some portion of total cost. In addition, it was felt that certain of the potential exogenous variables (gross ton miles of cars and contents -- GTMC -- and car miles, for example) would be highly correlated; i.e., ton miles necessitate car miles.

A preliminary correlation analysis (Table 2) revealed that such a situation did indeed exist. Most of distance related or weight and distance related output variables were very highly correlated as were the time variables with each other. To identify appropriate exogenous variables, therefore, stepwise regression procedures were used in conjunction with operations theory to specify the aggregate model.

Stepwise Regression

All of the output variables shown in Table 1 were included in a stepwise regression procedure with total cost. The results are depicted in Table 3.

As Table 3 indicates, the output (independent) variable most closely associated with total cost (TOTAL), car miles (CM), was brought into the equation first. The resulting overall "F test" for model appropriateness was highly significant.[1] The standard error of the estimate, the square root of the variance about the regression line, was relatively small, indicating fairly precise estimates. An R^2 of .97, furthermore, indicated that this variable alone explained 97% of the total variation in the dependent variable, TOTAL.

---

[1]The testing function is the ratio $\frac{\text{MS Regression}}{\text{MS Error}}$, which has an F distribution with K, n-K-1 degrees of freedom. The "P value" or probability of obtaining a greater F value is .0001.

TABLE 2. CORRELATIONS BETWEEN DEPENDENT AND INDEPENDENT VARIABLES.

	Total Railway Operating Expenditures	Tons of Freight Originated	Gross Ton Miles of Cars and Contents	Car Miles Running	Road Locomotive Unit Miles	Train Miles Running	Road Train Hours	Hours Yard Switching
Total	1.000	.938	.928	.943	.918	.934	.967	.972
Tons	.938	1.000	.890	.900	.864	.873	.933	.929
GTMC	.928	.890	1.000	.998	.981	.973	.833	.845
CM	.943	.900	.998	1.000	.984	.979	.855	.866
LUM	.918	.864	.981	.984	1.000	.979	.820	.824
TM	.934	.873	.973	.979	.979	1.000	.849	.860
ROADHR	.967	.933	.833	.855	.820	.840	1.000	.967
YARDHR	.972	.929	.845	.866	.824	.860	.967	1.000

TABLE 3. STEPWISE REGRESSION RESULTS FOR POTENTIAL INDEPENDENT VARIABLES.*

FORWARD SELECTION PROCEDURE FOR DEPENDENT VARIABLE TOTAL

STEP 1 VARIABLE CM ENTERED R SQUARE = 0.97282688 C(P) = 230.75716367

	DF	SUM OF SQUARES	MEAN SQUARE	F	PROB>F
REGRESSION	1	2899403495237097000.0000	2.899403495237E+19	1933.26	0.0001
ERROR	54	80986479539441740 0.0000	1.499749621101E+16		
TOTAL	55	2980389974776539100 0.0000			

	B VALUE	STD ERROR	TYPE II SS	F	PROB>F
INTERCEPT	7921919 9.66635149				
CM	0.75236401	0.01711130	2.899403495237E+19	1933.26	0.0001

STEP 2 VARIABLE YARDHR ENTERED R SQUARE = 0.99334279 C(P) = 19.27336299

	DF	SUM OF SQUARES	MEAN SQUARE	F	PROB>F
REGRESSION	2	2960548900824654200 0.0000	1.480274450412E+19	3954.15	0.0001
ERROR	53	198410739518847390.0000	3.743598858846E+15		
TOTAL	55	2980389974776539100 0.0000			

	B VALUE	STD ERROR	TYPE II SS	F	PROB>F
INTERCEPT	8646967.53545155				
CM	0.51955648	0.02012261	2.495664843384E+18	666.65	0.0001
YARDHR	414.23209497	32.41204867	6.114540559756E+17	163.33	0.0001

STEP 3 VARIABLE GTMC ENTERED R SQUARE = 0.99496694 C(P) = 4.37288992

	DF	SUM OF SQUARES	MEAN SQUARE	F	PROB>F
REGRESSION	3	2965389484876024600 0.0000	9.884631616253E+18	3426.56	0.0001
ERROR	52	150004899005143900.0000	2.884709596253E+15		
TOTAL	55	2980389974776539100 0.0000			

	B VALUE	STD ERROR	TYPE II SS	F	PROB>F
INTERCEPT	9847474.54608339				
CM	1.11304887	0.14595579	167759805499851080	58.15	0.0001
GTMC	-0.00853482	0.00208352	48405840513703472	16.79	0.0001
YARDHR	364.50363760	30.93361223	400533768836639630	138.85	0.0001

STEP 4 VARIABLE LUM ENTERED R SQUARE = 0.99533583 C(P) = 2.53429836

	DF	SUM OF SQUARES	MEAN SQUARE	F	PROB>F
REGRESSION	4	2966488923039983200 0.0000	7.416222307600E+18	2720.85	0.0001
ERROR	51	139010517365557230.0000	2.725696418932E+15		
TOTAL	55	2980389974776539100 0.0000			

	B VALUE	STD ERROR	TYPE II SS	F	PROB>F
INTERCEPT	6346964.21250206				
CM	0.90082630	0.17690281	70678954070641920	25.93	0.0001
GTMC	-0.00690654	0.00218152	27319840170647740	10.02	0.0026
LUM	1.88164531	0.93689539	10994381639586648	4.03	0.0499
YARDHR	395.25356808	33.74258714	374000622939952640	137.21	0.0001

NO OTHER VARIABLES MET THE 0.5000 SIGNIFICANCE LEVEL FOR ENTRY INTO THE MODEL.

*Does not include data for Conrail, which was excluded from the sample, as will be explained in a later section of the study. Includes only those railroads which provide unit train service.

Constructing a hypothesis test for the model in Step One (Table 4) it becomes apparent that the simple linear model of car miles on total cost is a good approximation of the relationship between cost and output.

In Step Two of the procedure, the variable "yard switching hours" (YARDHR) was brought into the equation. As Table 3 indicates, addition of the variable caused a significant improvement in the explanatory value of the model, as measured by the partial F statistic of 163; highly significant at the 99% confidence level.[2] The R^2 also increased as a greater proportion of the variance of Y was explained by the expanded model.

In Step Three, the variable "gross ton miles of cars and contents" (GTMC) was brought into the regression. As before, the addition of the new variable increased the proportion of the variance in TOTAL explained by the model. The partial F test was significant and the R^2 increased slightly. However, at this stage of the procedure, a problem occurred. The variable GTMC had an unexpected (negative) sign. This is contrary to operations logic. Furthermore, as noted in Table 1, the variables CM and GTMC are highly correlated. The negative sign, therefore, may well be a sign of multicollinearity. For these reasons, a decision

---

[2]The partial F denotes the significance of the extra or incremental reduction in the unexplained portion of the sum of squares of TOTAl caused by the addition by YARDHR to the model. The general formula for the incremental SS for a two variable model is given by:

$$SS(X_2|X_1) = SSR\ (X_1,X_2) - SSR\ (X_1).$$

The test function is the ratio $SS(X_2|X_1)/MSE\ (X_1,X_2)$. This ratio is F distributed with 1 and (n-p-2) degrees of freedom under the H_0.

TABLE 4. HYPOTHESIS TEST: STAGE ONE OF STEPWISE REGRESSION

Item	Description/Value
Testing Function	Mean square regression/mean square error
Test Statistic	F with K, n-K-1 degrees of freedom
H_o:	There is no linear relationship between CM and TOTAL; $B_1 = 0$.
Decision Rule:	If F_{cal} 7 F, n, n-k-1, for an alpha of .01, then reject H_o
Conclusion:	Reject H_o

was made to halt the stepwise procedure after the second stage.[4]

The aggregate model which thus came out of the stepwise procedure was:

(1) $\hat{Y} = \hat{B}_0 + \hat{B}_1 X_1 + \hat{B}_2 X_2 + E$

where:

X_1 = car miles

X_2 = yard switching hours

E = error term

$\hat{B}$ = estimated coefficients

This model, however, it will be noted, says nothing about the effects of unit train output on costs. After the major causal variables had been identified, therefore, unit train measures were introduced into the equation in an effort to explain the effects of unit train output on cost.

[4]While an additional variable, road locomotive unit miles (LUM), was added after GTMC, the variable was barely significant at the 95% confidence level. Therefore, because of this and because of its high correlation to car miles, it was decided not to include the variable in the model.

Unit Train Model

The model depicted in equation (1) was thus modified to account for unit train output, as depicted below:

$$(2)\ \hat{Y} = \hat{B}_0 + \hat{B}_1X_1 + \hat{B}_2X_2 + \hat{B}_3X_3 + E$$

where:

X_3 = unit train miles of output.

The sample used, as noted earlier, contains only those railroads which originated unit train traffic in 1979 or 1980.

Table 5 depicts the results of the respecified model. Several things will be discussed on the basis of this and subsequent tables. First, the adequacy of the model containing a unit train output variable will be noted. Second, an analysis of the residuals of the regression will be undertaken. And third, the question of multicollinearity will be addressed.

IV. INTERPRETATION OF STATISTICAL RESULTS

First of all, a unit train model calibrated on the basis of Class I railroads which originate unit train traffic (including Conrail) is clearly a significant aid in explaining the variance of total cost. The overall F test is significant at the 99% confidence level; an R^2 of nearly .99 indicates that 99% of the variation in TOTAL is explained by the model; and a coefficient of variation $[(\sigma/\bar{Y})\ 100]$ of 13.6 indicates that while there is considerable variation about the dependent variable mean, this does not appear unduly troublesome given the wide range of railroad sizes and configurations.

TABLE 5. RESULTS OF REGRESSION ANALYSIS FOR UNIT TRAIN COST MODEL, INCLUDING RESIDUAL AND COLLINEARITY DIAGNOSTICS.

DEP VARIABLE: TOTAL

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	3	5.29128E+19	1.76376E+19	1578.548	0.0001
ERROR	54	6.03358E+17	1.11733E+16		
C TOTAL	57	5.35162E+19			

ROOT MSE	105703848	R-SQUARE	0.9887
DEP MEAN	774546707	ADJ R-SQ	0.9881
C.V.	13.64719		

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > |T|	TOLERANCE
INTERCEP	1	-52011066	18357817	-2.833	0.0065	.
CM	1	0.459354	0.043044	10.672	0.0001	0.098021
UTM	1	-24.752948	10.044555	-2.464	0.0169	0.327429
YARDHR	1	650.420	40.604378	16.018	0.0001	0.163921

OBS	_ID	ACTUAL	PREDICT VALUE	STD ERR PREDICT	RESIDUAL	STD ERR RESIDUAL	STUDENT RESIDUAL	-2-1-0 1 2	COOK'S D
1	BO	1.0E+09	1.1E+09	23728937	-1.3E+08	1.0E+08	-1.311	**	0.023
2	BO	8.8E+08	1.0E+09	19098170	-1.2E+08	1.0E+08	-1.122	**	0.011
3	BM	1.3E+08	83633586	17288190	48560414	1.0E+08	0.466		0.001
4	BM	1.1E+08	71090416	17358780	40313584	1.0E+08	0.387		0.001
5	CO	8.7E+08	9.1E+08	17304579	-3.4E+07	1.0E+08	-0.329		0.001
6	CO	8.3E+08	9.8E+08	17228964	-1.5E+08	1.0E+08	-1.450	**	0.014
7	CR	4.6E+09	4.3E+09	74630415	3.2E+08	74857228	4.276	******	4.543
8	CR	3.7E+09	3.5E+09	51701703	1.7E+08	92196732	1.851	***	0.269
9	DH	1.3E+08	40203579	17717303	90303421	1.0E+08	0.867	*	0.005
10	DH	1.2E+08	37718445	17764407	80986555	1.0E+08	0.777	*	0.004
11	DTI	96029000	37407439	17631780	58621561	1.0E+08	0.562	*	0.002
12	DTI	76586000	13321101	17790668	63264899	1.0E+08	0.607	*	0.003
13	EJE	1.3E+08	1.9E+08	21366851	-6.4E+07	1.0E+08	-0.621	*	0.004
14	EJE	93796000	1.1E+08	19146903	-1.4E+07	1.0E+08	-0.138		0.000
15	GTW	2.4E+08	3.2E+08	19209291	-7.1E+07	1.0E+08	-0.687	*	0.004
16	GTW	2.0E+08	2.3E+08	17808692	-3.2E+07	1.0E+08	-0.310		0.001
17	PLE	59495000	1.2E+08	19074060	-6.5E+07	1.0E+08	-0.627	*	0.003
18	PLE	62890000	88602287	18631815	-2.6E+07	1.0E+08	-0.247		0.000
19	WM	65745000	18361644	17763540	47383356	1.0E+08	0.455		0.002
20	WM	84308000	15155159	17765542	69152841	1.0E+08	0.664	*	0.003
21	AGS	92735000	73471312	17511365	19263688	1.0E+08	0.185		0.000
22	AGS	85600000	64153717	17516419	21446283	1.0E+08	0.206		0.000
23	CGA	1.2E+08	99100410	17006493	24139590	1.0E+08	0.231		0.000
24	CGA	1.2E+08	1.1E+08	16956047	19492157	1.0E+08	0.187		0.000
25	CNTP	1.5E+08	1.3E+08	17779427	18288284	1.0E+08	0.176		0.000
26	CNTP	1.3E+08	1.1E+08	17680296	14405780	1.0E+08	0.138		0.000
27	ICG	1.0E+09	1.1E+09	18355446	-5.3E+07	1.0E+08	-0.512	*	0.002
28	ICG	9.5E+08	9.6E+08	16584452	-1.8E+07	1.0E+08	-0.172		0.000
29	LN	1.1E+09	1.2E+09	18327918	-1.1E+08	1.0E+08	-1.028	**	0.009
30	LN	1.0E+09	1.1E+09	20042990	-9.0E+07	1.0E+08	-0.870	*	0.007
31	SOU	9.5E+08	9.4E+08	17262861	18676460	1.0E+08	0.179		0.000
32	SOU	9.0E+08	9.1E+08	15791406	-1.0E+07	1.0E+08	-0.097		0.000
33	ATSF	2.0E+09	1.9E+09	32607444	55991289	1.0E+08	0.557	*	0.008
34	ATSF	2.0E+09	1.9E+09	32564453	52631157	1.0E+08	0.523	*	0.007
35	BN	2.7E+09	2.8E+09	66222700	-6.3E+07	82388455	-0.759	*	0.093
36	BN	2.7E+09	2.7E+09	78175695	8828835	71146779	0.124		0.005
37	CNW	8.6E+08	1.0E+09	17095448	-1.6E+08	1.0E+08	-1.497	**	0.015
38	CNW	8.7E+08	1.0E+09	15401465	-1.4E+08	1.0E+08	-1.353	**	0.010
39	MILW	6.6E+08	6.1E+08	15374611	40426816	1.0E+08	0.387		0.001
40	MILW	4.6E+08	4.4E+08	15572521	24570740	1.0E+08	0.235		0.000
41	CS	1.2E+08	9679776	18305159	1.1E+08	1.0E+08	1.102	**	0.009
42	CS	1.2E+08	16291401	18837609	1.1E+08	1.0E+08	1.011	**	0.008
43	DRGW	2.5E+08	2.1E+08	16412694	41201914	1.0E+08	0.395		0.001
44	DRGW	2.3E+08	1.9E+08	16826124	45400444	1.0E+08	0.435		0.001
45	DMIR	91640000	32442374	17595255	59197626	1.0E+08	0.568	*	0.002
46	DMIR	73732000	14242051	17743886	59489949	1.0E+08	0.571	*	0.002
47	FWD	93211000	40281060	17980730	52929940	1.0E+08	0.508	*	0.002
48	FWD	1.2E+08	61912176	18564240	55712824	1.0E+08	0.535	*	0.002
49	KCS	2.4E+08	2.6E+08	16493727	-2.6E+07	1.0E+08	-0.253		0.000
50	KCS	2.4E+08	2.7E+08	16446277	-2.8E+07	1.0E+08	-0.264		0.000
51	MKT	2.0E+08	1.9E+08	16350637	7133527	1.0E+08	0.068		0.000
52	MKT	2.1E+08	2.2E+08	16112477	-7408897	1.0E+08	-0.071		0.000
53	MP	1.5E+09	1.6E+09	19655859	-1.3E+08	1.0E+08	-1.250	**	0.014
54	MP	1.5E+09	1.6E+09	20664019	-1.3E+08	1.0E+08	-1.287	**	0.016
55	SP	2.2E+09	2.1E+09	49611820	81159657	93337938	0.870	*	0.053
56	SP	2.0E+09	2.4E+09	34885661	-4.3E+08	99731231	-4.295	******	0.564
57	UP	1.8E+09	1.7E+09	50138799	48040089	93055920	0.516	*	0.019
58	UP	1.7E+09	1.6E+09	47573098	1.2E+08	94393346	1.323	**	0.111

SUM OF RESIDUALS .00000482798
SUM OF SQUARED RESIDUALS 6.03358E+17

The tests for significance of the individual parameters shown in Table 5 (the T test for the null hypothesis that $\hat{B} = 0$) are all significant at the 95% level. The sign of the parameters, furthermore, is logical as well. Here, it is expected that unit train miles would have a negative sign; that is, after controlling for CM and YARDHR the effect of unit train miles is to lower cost. For example, every yard hour incurred switching results in a cost of $650, while an expense of 46¢ is incurred per car mile. However, if the shipment is a unit train, a cost savings of $24 per mile will be incurred over non-unit train shipments.

Residual Analysis

Table 5 also presents detailed data concerning the residuals of the regression, or the portion of the variation in Y ($Y_i - \bar{Y}$) which is not explained by the straight-line regression.

An analysis of outliers immediately flagged two railroads: Conrail, or Consolidation Railway Corporation, and the Southern Pacific. It was anticipated that Conrail would be an outlier, because of poor financial history and the current situation of the railroad.[5] Conrail's 1979 standardized residual value lay 3 standard deviations from the mean (the mean of all E_i's is zero). On the basis of an understanding of the history and financial posture of the railroad, and on the basis of a large residual value, it was decided to eliminate Conrail from the sample.

[5] Conrail is a government-created and financed amalgamation of several bankrupted railroads in the northeastern U.S. The carrier started with a poor physical system, little or no financial capital, and stiff truck competition for its traffic base. It was therefore anticipated that a large positive difference would exist between the actual and predicted values.

The second extreme case, the Southern Pacific, was somewhat more difficult to evaluate. It was not expected that this railroad would exhibit a negative deviation of 3 SD's about the mean in terms of its residual value. Furthermore, in the second year of the data, the Southern Pacific exhibited a small, "near-normal" deviation of 1 SD about the mean. It was decided, therefore, to rerun the analysis after excluding Conrail prior to making a final determination regarding the SP.

As Table 6 denotes, excluding Conrail from the sample did nothing but good things for the model. The fit improved dramatically for the exclusion of just one case. The F statistic more than doubled. The sum of squares of "total" was reduced by over 1/3. The coefficient of variation, consequently, dropped to 7.73. The parameter estimates themselves change (CM increasing to .60, YARDHR dropping to 357 and UTM remaining relatively stable). All statistical tests for significance (T-tests) were now significant at the 99% level, furthermore. The SP, however, remained a severe outlier but now for both years, and in opposite directions. It was decided, therefore, to eliminate the Southern Pacific from the final analysis as well.

Table 7 depicts of the results of the regression analysis for the reformulated data set excluding the SP. As was the case with the exclusion of Conrail, the omission of SP from the data base served to improve the fit of the linear model. Once again, a further reduction in the unexplained sum of squares for "total" was achieved. The F statistic increased by 42%, the R^2 increased slightly, and the coefficient of

TABLE 6. RESULTS OF REGRESSION ANALYSIS FOR UNIT TRAIN COST MODEL EXCLUDING CONRAIL DATA.

DEP VARIABLE: TOTAL

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	3	2.96706E+19	9.89021E+18	3858.675	0.0001
ERROR	52	1.33282E+17	2.56311E+15		
C TOTAL	55	2.98039E+19			

ROOT MSE	50627160	R-SQUARE	0.9955
DEP MEAN	654864286	ADJ R-SQ	0.9953
C.V.	7.730939		

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > |T|	TOLERANCE
INTERCEP	1	6608467	9803253	0.674	0.5032	.
CM	1	0.600853	0.023180	25.921	0.0001	0.093126
UTM	1	-24.384546	4.837387	-5.041	0.0001	0.324099
YARDHR	1	357.212	29.107080	12.272	0.0001	0.153236

OBS	ID	ACTUAL	PREDICT VALUE	STD ERR PREDICT	RESIDUAL	STD ERR RESIDUAL	STUDENT RESIDUAL	-2-1-0 1 2	COOK'S D
1	BO	1.0E+09	9.5E+08	18424683	61274097	47155492	1.299	|**	0.064
2	BO	8.8E+08	8.6E+08	14021445	27048315	48646772	0.556	|*	0.006
3	BM	1.3E+08	1.1E+08	8556191	18929913	49897191	0.379	|	0.001
4	BM	1.1E+08	1.0E+08	8655394	7326768	49880059	0.147	|	0.000
5	CO	8.7E+08	7.9E+08	11933596	81891832	49199383	1.664	|***	0.041
6	CO	8.3E+08	8.6E+08	12186896	-3.0E+07	49138467	-0.611	*|	0.006
7	DH	1.3E+08	1.0E+08	9569732	30397285	49714481	0.611	|*	0.003
8	DH	1.2E+08	99164959	9642736	19540041	49700372	0.393	|	0.001
9	DTI	96029000	76223560	8920504	19805440	49835063	0.397	|	0.001
10	DTI	76586000	59720716	9186673	16865284	49786689	0.339	|	0.001
11	EJE	1.3E+08	1.5E+08	10754628	-2.0E+07	49471682	-0.402	|	0.002
12	EJE	93796000	1.0E+08	9195417	-6516568	49785075	-0.131	|	0.000
13	GTW	2.4E+08	2.7E+08	9911436	-2.2E+07	49647485	-0.434	|	0.002
14	GTW	2.0E+08	2.1E+08	8649240	-1.3E+07	49882863	-0.260	|	0.001
15	PLE	59495000	1.2E+08	9169103	-5.6E+07	49789929	-1.118	**|	0.011
16	PLE	62890000	92736133	8935230	-3.0E+07	49832429	-0.599	*|	0.003
17	WM	65745000	61221132	9079995	4523868	49806255	0.091	|	0.000
18	WM	84308000	60012628	9133089	24295372	49796546	0.488	|	0.002
19	AGS	92735000	1.3E+08	9479629	-4.1E+07	49731740	-0.816	*|	0.005
20	AGS	85600000	1.2E+08	9493071	-3.9E+07	49729176	-0.778	*|	0.006
21	CGA	1.2E+08	1.4E+08	8662707	-1.6E+07	49880526	-0.316	|	0.001
22	CGA	1.2E+08	1.4E+08	8601297	-1.9E+07	49891152	-0.378	|	0.001
23	CNTP	1.5E+08	1.9E+08	9758257	-4.6E+07	49677820	-0.931	*|	0.008
24	CNTP	1.3E+08	1.8E+08	9691968	-4.9E+07	49690795	-0.995	*|	0.009
25	ICG	1.0E+09	9.3E+08	12603633	68997640	49033232	1.407	|**	0.033
26	ICG	9.5E+08	8.7E+08	10665499	78372473	49490974	1.584	|***	0.029
27	LN	1.1E+09	1.1E+09	11662646	-4234831	49265526	-0.086	|	0.000
28	LN	1.0E+09	1.0E+09	10419497	-3.8E+07	49543349	-0.769	*|	0.007
29	SOU	9.5E+08	9.3E+08	8307729	20031813	49940875	0.401	|	0.001
30	SOU	9.0E+08	9.1E+08	7590409	-7647700	50054920	-0.153	|	0.000
31	ATSF	2.0E+09	1.9E+09	15837872	81519521	480960 81	1.695	|***	0.078
32	ATSF	2.0E+09	1.9E+09	15740285	68747375	481181 13	1.429	|**	0.055
33	BN	2.7E+09	2.7E+09	32191534	11068014	390744 74	0.283	|	0.014
34	BN	2.7E+09	2.7E+09	37548114	-2.4E+07	339595 12	-0.719	*|	0.158
35	CNW	8.6E+08	8.9E+08	12179286	-3.4E+07	491403 54	-0.699	*|	0.007
36	CNW	8.7E+08	9.1E+08	10247513	-4.5E+07	495792 08	-0.918	*|	0.009
37	MILW	6.6E+08	5.7E+08	8177321	88565872	499623 93	1.773	|***	0.021
38	MILW	4.6E+08	4.2E+08	7579832	42858755	500565 15	0.856	|*	0.004
39	CS	1.2E+08	80798617	10226171	43587383	495836 14	0.879	|*	0.008
40	CS	1.2E+08	92460222	10641571	28996778	494061 24	0.586	|*	0.004
41	DRGW	2.5E+08	2.5E+08	8397698	1266618	499253 25	0.025	|	0.000
42	DRGW	2.3E+08	2.3E+08	8683909	1729210	498768 39	0.035	|	0.000
43	DMIR	91640000	76284854	9029281	15355146	498154 74	0.308	|	0.001
44	DMIR	73732000	63282903	9239432	10449097	497769 25	0.210	|	0.000
45	FWD	93211000	98540299	9633226	-5329299	497022 16	-0.107	|	0.000
46	FWD	1.2E+08	1.2E+08	9924660	-3662463	49644 43	-0.074	|	0.000
47	KCS	2.4E+08	2.6E+08	7902088	-2.4E+07	500066 63	-0.489	|	0.001
48	KCS	2.4E+08	2.7E+08	7880363	-3.0E+07	500100 91	-0.593	*|	0.002
49	MKT	2.0E+08	2.1E+08	8001489	-1.5E+07	499903 54	-0.302	|	0.001
50	MKT	2.1E+08	2.4E+08	7880756	-2.9E+07	500100 29	-0.581	*|	0.002
51	MP	1.5E+09	1.5E+09	11830613	-3.3E+07	492134 17	-0.674	*|	0.007
52	MP	1.5E+09	1.5E+09	11025036	-6.9E+07	494121 23	-1.404	**|	0.025
53	SP	2.2E+09	2.1E+09	24345473	1.4E+08	443892 270	3.121	|******	0.732
54	SP	2.0E+09	2.2E+09	25011684	-1.8E+08	440173 326	-4.064	******|	1.333
55	UP	1.8E+09	1.8E+09	25421694	-6.0E+07	437813 309	-1.378	**|	0.160
56	UP	1.7E+09	1.8E+09	25318217	-2.3E+07	438417 29	-0.517	*|	0.022

SUM OF RESIDUALS 7.45058E-08
SUM OF SQUARED RESIDUALS 1.33282E+17

TABLE 7. RESULTS OF REGRESSION ANALYSIS FOR UNIT TRAIN COST MODEL EXCLUDING BOTH CONRAIL AND SOUTHERN PACIFIC.

DEP VARIABLE: TOTAL

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	3	2.53954E+19	8.46515E+18	5498.022	0.0001
ERROR	50	7.69836E+16	1.53967E+15		
C TOTAL	53	2.54724E+19			

ROOT MSE	39238650	R-SQUARE	0.9970
DEP MEAN	601489500	ADJ R-SQ	0.9968
C.V.	6.52358		

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > |T|	TOLERANCE
INTERCEP	1	-74304.488	7870743	-0.009	0.9925	.
CM	1	0.576659	0.019272	29.922	0.0001	0.090868
UTM	1	-23.311965	4.299534	-5.422	0.0001	0.247131
YARDHR	1	399.368	23.654300	16.884	0.0001	0.170321

OBS	ID	ACTUAL	PREDICT VALUE	STD ERR PREDICT	RESIDUAL	STD ERR RESIDUAL	STUDENT RESIDUAL	-2-1-0	1 2	COOK'S D
1	BO	1.0E+09	9.8E+08	15611414	34319613	35939381	0.553		*	0.043
2	BO	8.8E+08	8.7E+08	12035490	7543560	37347271	0.202			0.001
3	BM	1.3E+08	1.1E+08	6745249	21749969	38654537	0.563		*	0.002
4	BM	1.1E+08	1.0E+08	6837877	10598311	38638259	0.274			0.001
5	CO	8.7E+08	8.1E+08	10218761	66090785	37834675	1.745		***	0.055
6	CO	8.3E+08	8.8E+08	10580116	-4.6E+07	37785352	-1.224	**		0.029
7	DH	1.3E+08	92792294	7580820	37714706	38499337	0.980		*	0.009
8	DH	1.2E+08	91630561	7644181	27074439	38436857	0.703		*	0.005
9	DTI	96029000	72220431	7079455	23808569	38594727	0.617		*	0.003
10	DTI	76586000	54684833	7325521	21901167	38548779	0.568		*	0.003
11	EJE	1.3E+08	1.6E+08	9526635	-2.9E+07	38301020	-0.730	*		0.007
12	EJE	93796000	1.0E+08	7271868	-9363146	38558937	-0.243			0.001
13	GTW	2.4E+08	2.7E+08	7808970	-3.0E+07	38453760	-0.777	*		0.006
14	GTW	2.0E+08	2.1E+08	6758683	-1.7E+07	38652191	-0.440			0.001
15	PLE	59495000	1.2E+08	7248343	-5.9E+07	38553366	-1.523	***		0.020
16	PLE	62890000	93870007	7079735	-3.1E+07	38594676	-0.803	*		0.005
17	WM	65745000	56660419	7219549	9084581	38568767	0.236			0.000
18	WM	84308000	55145646	7260276	29162354	38561121	0.756		*	0.005
19	AGS	92735000	1.3E+08	7498205	-3.3E+07	38515563	-0.861	*		0.007
20	AGS	85600000	1.2E+08	7524206	-3.1E+07	38510492	-0.814	*		0.006
21	CGA	1.2E+08	1.3E+08	6815172	-1.1E+07	38642271	-0.293			0.001
22	CGA	1.2E+08	1.4E+08	6769831	-1.5E+07	38650240	-0.379			0.001
23	CNTP	1.5E+08	1.8E+08	7606758	-3.8E+07	38476377	-0.983	*		0.010
24	CNTP	1.3E+08	1.7E+08	7648069	-4.1E+07	38436034	-1.071	**		0.011
25	ICG	1.0E+09	9.5E+08	10435442	51907682	37811456	1.373		**	0.036
26	ICG	9.5E+08	8.8E+08	8843186	64871651	38229174	1.697		***	0.039
27	LN	1.1E+09	1.1E+09	11065499	-1.7E+07	37646067	-0.439			0.004
28	LN	1.0E+09	1.0E+09	9932033	-4.3E+07	37947736	-1.136	**		0.022
29	SOU	9.5E+08	9.3E+08	7276763	21158900	38553913	0.549		*	0.003
30	SOU	9.0E+08	9.1E+08	6435608	-7112027	38707294	-0.184			0.000
31	ATSF	2.0E+09	1.9E+09	14669251	82135750	36393874	2.257		****	0.207
32	ATSF	2.0E+09	1.9E+09	14360264	70501785	36516496	1.931		***	0.144
33	BN	2.7E+09	2.7E+09	25146442	-488951	30121888	-0.016			0.000
34	BN	2.7E+09	2.7E+09	29697199	-2.2E+07	25646598	-0.860	*		0.248
35	CNW	8.6E+08	9.1E+08	10602443	-5.1E+07	37779093	-1.340	**		0.035
36	CNW	8.7E+08	9.3E+08	8993752	-5.8E+07	38194032	-1.522	***		0.032
37	MILW	6.6E+08	5.7E+08	6522412	81312960	38692762	2.102		****	0.031
38	MILW	4.6E+08	4.2E+08	5902461	39417019	38792172	1.016		**	0.006
39	CS	1.2E+08	72569351	8332059	51816649	38332920	1.352		**	0.022
40	CS	1.2E+08	83717546	8832181	37739454	38231718	0.987		*	0.013
41	DRGW	2.5E+08	2.4E+08	6661420	5614597	38669072	0.145			0.000
42	DRGW	2.3E+08	2.3E+08	6953805	6423874	38616663	0.166			0.000
43	DMIR	91640000	71472004	7151212	20167996	38581496	0.523		*	0.002
44	DMIR	73732000	57756689	7338718	15975311	38546269	0.414			0.002
45	FWD	93211000	92079802	7853649	1131198	38444659	0.029			0.000
46	FWD	1.2E+08	1.1E+08	8208621	2704062	38370434	0.070			0.000
47	KCS	2.4E+08	2.6E+08	6152714	-2.6E+07	38753268	-0.669	*		0.003
48	KCS	2.4E+08	2.7E+08	6154044	-3.1E+07	38753057	-0.789	*		0.004
49	MKT	2.0E+08	2.1E+08	6221535	-1.3E+07	38742279	-0.333			0.001
50	MKT	2.1E+08	2.4E+08	6124596	-2.7E+07	38757722	-0.694	*		0.003
51	MP	1.5E+09	1.5E+09	11332533	-4.4E+07	37566546	-1.175	**		0.031
52	MP	1.5E+09	1.6E+09	10216203	-7.6E+07	37885364	-2.006	****		0.073
53	UP	1.3E+09	1.3E+09	21761722	-4.0E+07	32651173	-1.218	**		0.165
54	UP	1.7E+09	1.7E+09	20852994	2095196	33239898	0.063			0.000

SUM OF RESIDUALS .00000102818
SUM OF SQUARED RESIDUALS 7.69836E+16

variation declined to 6.5. And, in general, the value of the parametric test for each of the independent variables in the model improved. While there were still instances of large discrepancies between the observed and predicted values for several cases, nothing approached the extreme values generated by Conrail and SP in the raw data set.

After reformulation of the original data set to exclude the two extreme outliers, a visual inspection of the residual plots were made. Inspection of these plots, depicted in Tables 8-11, indicates general conformance with the assumptions underlying the linear model; i.e., homoscedasticity and independence of the E_i's associated with various combinations of values of the independent variables.

As Table 8 indicates, the residuals for the unit train model scatter in a fairly random pattern about the mean of E, indicating that the assumption of linearity for the overall model appears to be an appropriate supposition. The residuals, in addition, were plotted against each individual output variable, searching for signs of potential nonlinearity or violations of the independence assumptions. As Table 9 indicates, the plot of residuals against car miles is once again random in nature, as is the plot of residuals against the variable "yard switching hours", depicted in Table 10. Only in the case of unit train miles is there reason to perhaps ponder over the residual scatter (Table 11). The majority of the residuals scatter aimlessly at lower levels of output. There is considerable difference in the value of two observations for the data set, which appear at the extreme right

TABLE 8. PLOT OF RESIDUALS AGAINST TOTAL COST FOR UNIT TRAIN COST MODEL.

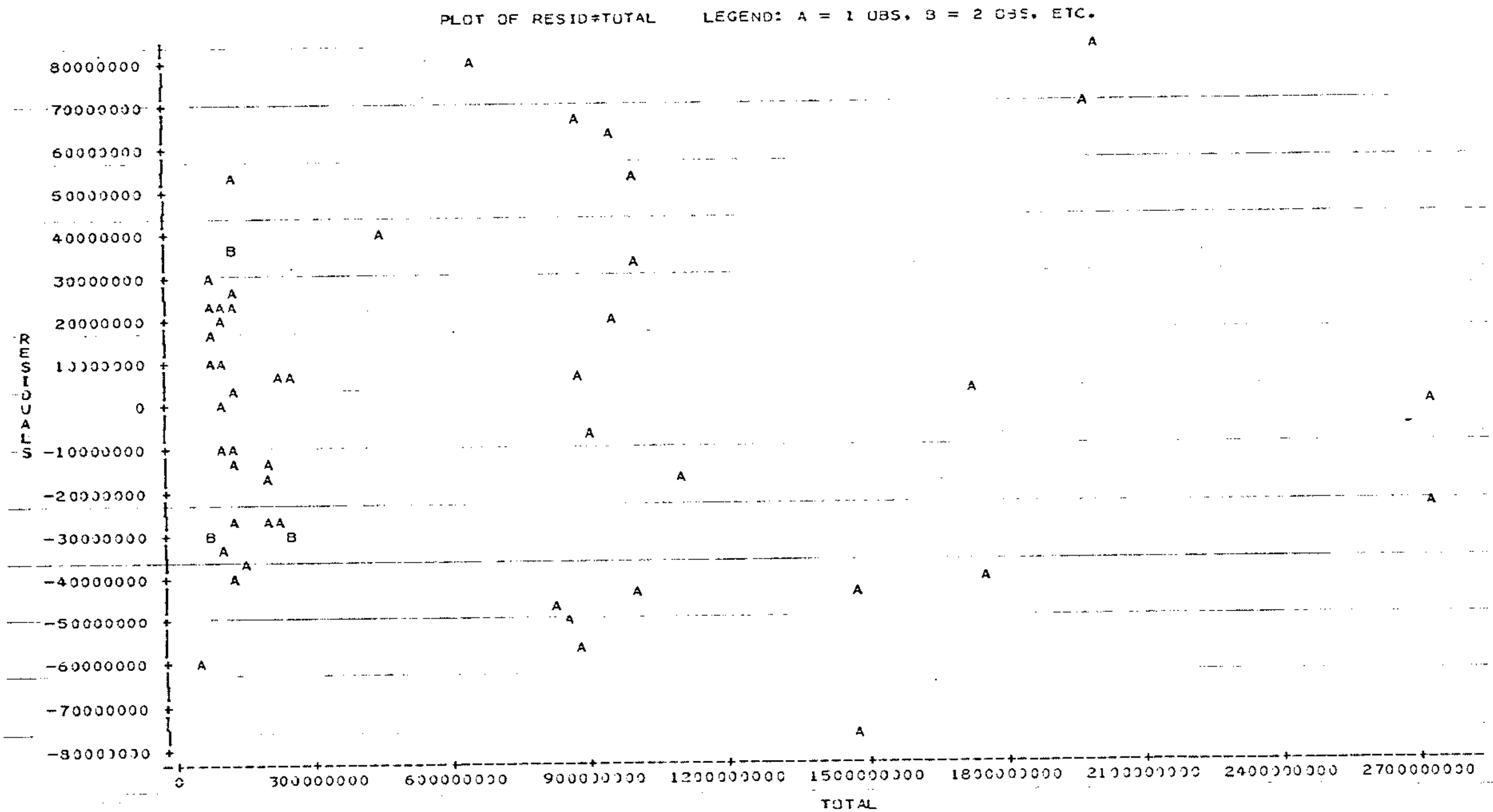


TABLE 9. PLOT OF RESIDUALS AGAINST CAR MILES FOR UNIT TRAIN COST MODEL.

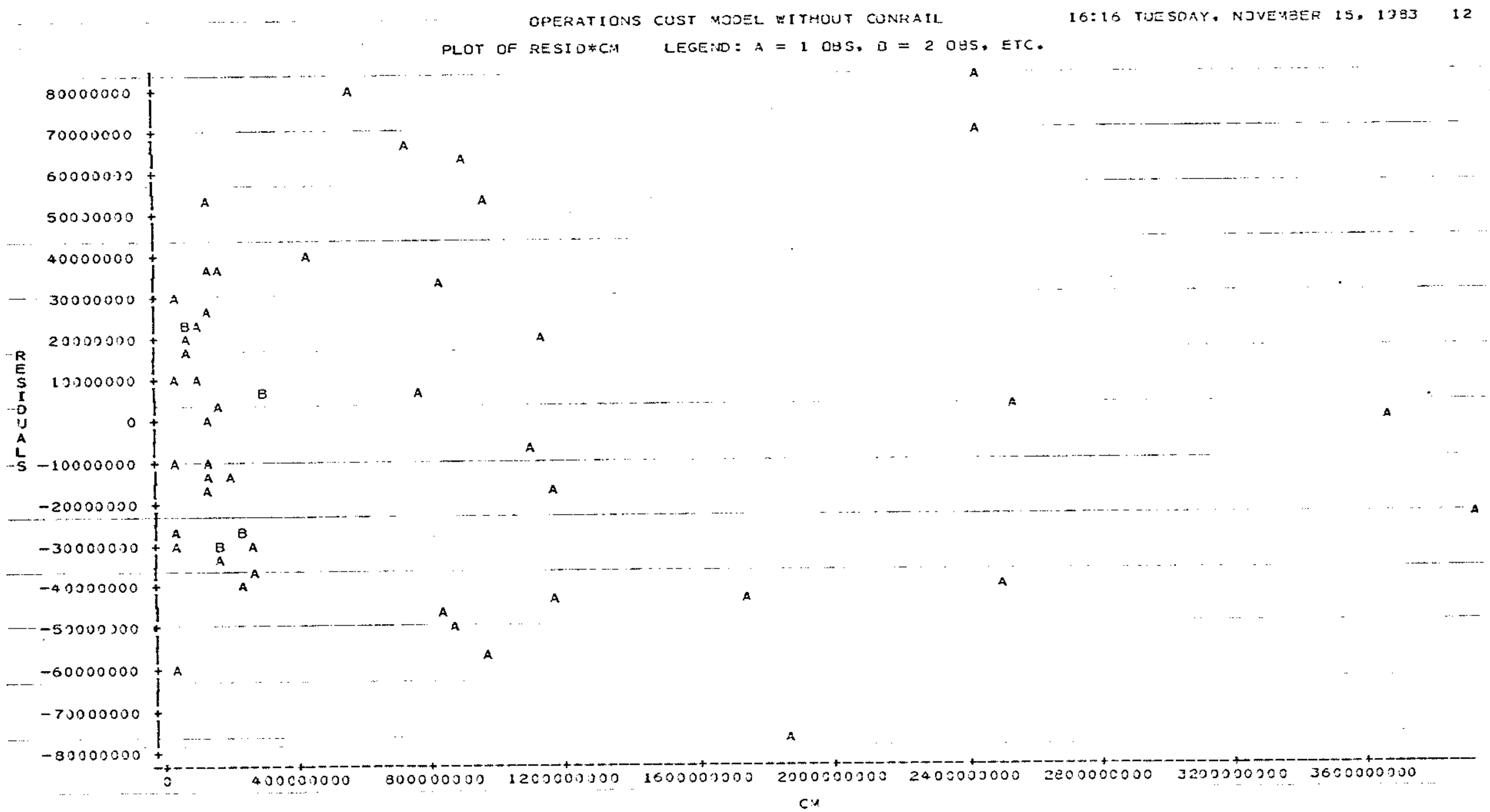


TABLE 10. PLOT OF RESIDUALS AGAINST YARD HOURS SWITCHING FOR UNIT TRAIN COST MODEL.

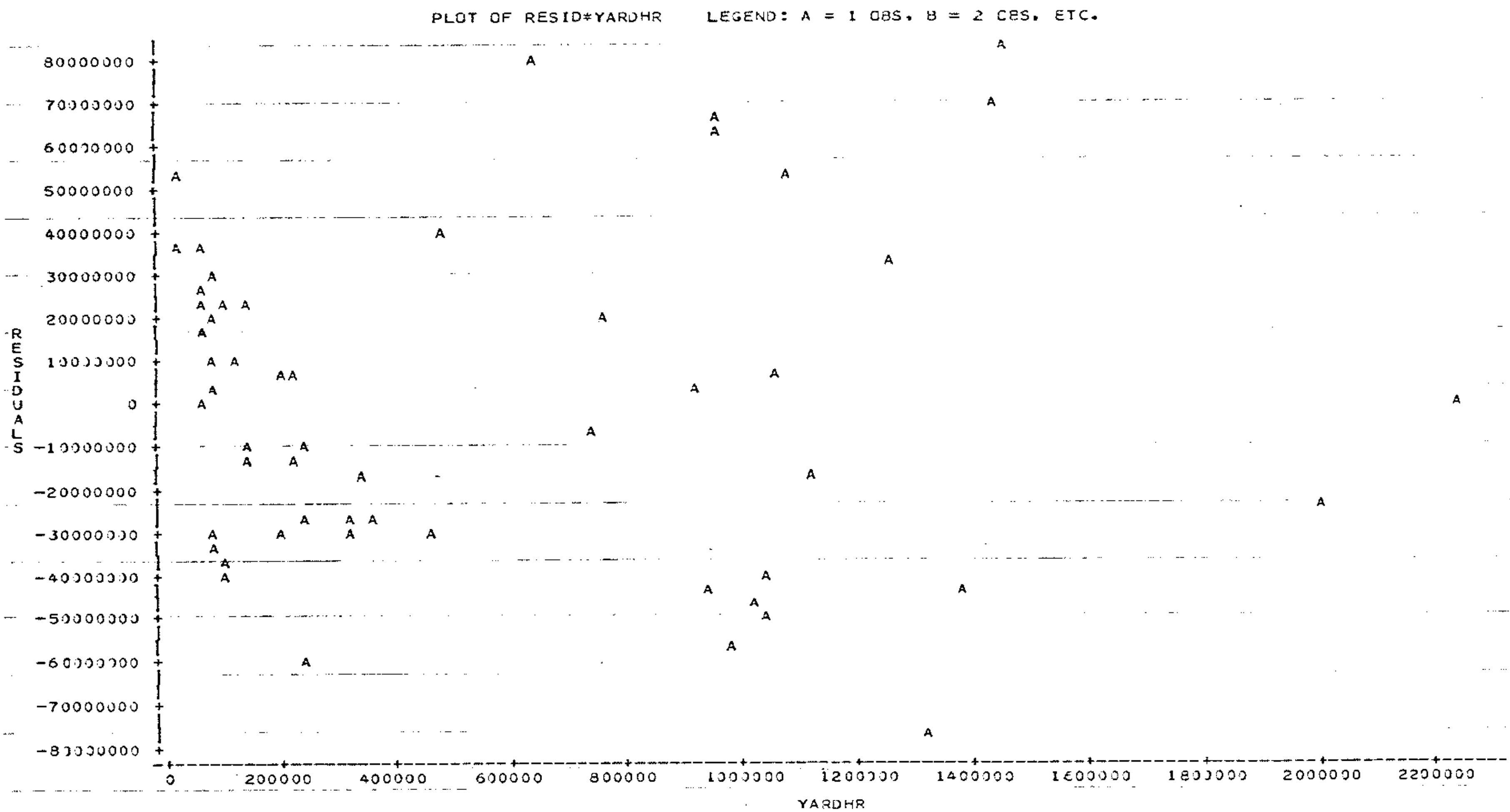
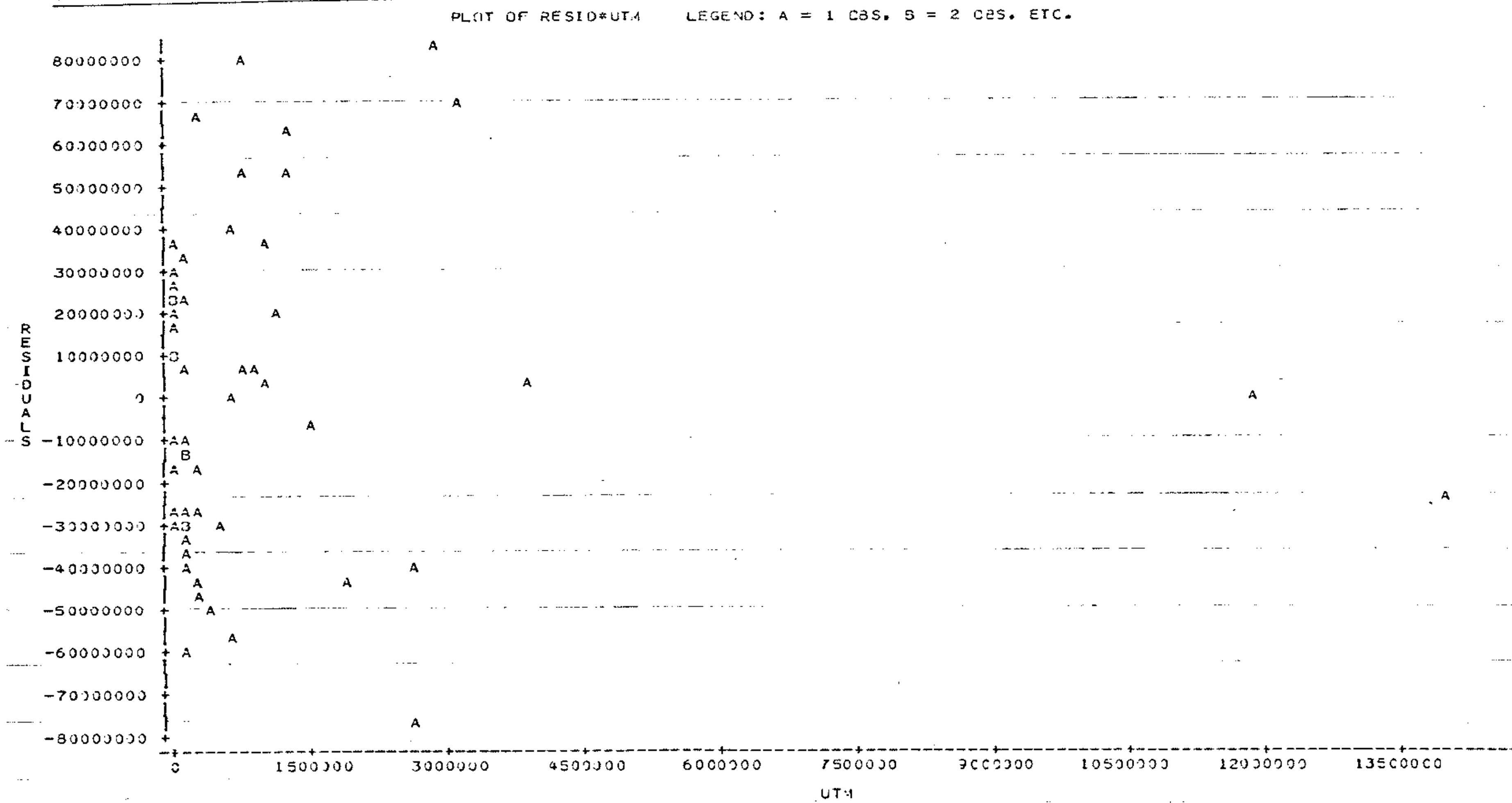


TABLE 11. PLOT OF RESIDUALS AGAINST UNIT TRAIN MILES OF OUTPUT.



of the quadrant (Burlington Northern for years 1979-1980). If these two observations are considered as a subpopulation, then their variance is clearly less than the remainder of the residuals, which may be defined as a second subpopulation. Any overall pattern, however, is weak at best.

As a general rule, the simple fact that some errors are larger than others is not, in and of itself, sufficient to establish a violation of homoscedasticity. Rather, "unless pattern is found, the convenient assumption of homoscedasticity should be accepted".[6] And so it will in this instance.

Test for Differences Between Years

A related problem with regard to using pooled data is whether or not there is a signficant difference across years after inflating costs for inflation. If so, it is desirable that this be controlled for by bringing a "dummy" variable into the equation.

To test for such possible differences, a dummy variable was created (1 if year = 1979, 0 otherwise) and the regression re-ran including the variable DYEAR. As Table 12 indicates, the introduction of the dummy variable added little to the explanatory value of the model. The T-test for the null hypothesis that $\hat{B}_3=0$ could not be rejected at any reasonable alpha. It was concluded, therefore, that the dummy should be dropped from the equation; the implication being that there really is no significant difference between the two years of data.

---

[6] Beals, Ralph E. Statistics for Economists, Rand McNally and Company, Chicago, Illinois, 1972: Chapter 13, "Distribution of Errors", page 343.

TABLE 13. ANOVA TABLE FOR MODEL INCLUDING DUMMY VARIABLE.

DEP VARIABLE: TOTAL

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	4	2.53965E+19	6.34913E+18	4097.812	0.0001
ERROR	49	7.59203E+16	1.54939E+15		
C TOTAL	53	2.54724E+19			

ROOT MSE	39362349	R-SQUARE	0.9970
DEP MEAN	601489500	ADJ R-SQ	0.9968
C.V.	6.544146		

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > |T|
INTERCEP	1	-4121123	9284628	-0.444	0.6591
UTM	1	-23.177160	4.316157	-5.370	0.0001
CM	1	0.577566	0.019364	29.827	0.0001
YARDHR	1	397.167	23.877094	16.634	0.0001
DYEAR	1	8955581	10810766	0.828	0.4115

Multicollinearity Diagnosis

A likely problem with multiple regression analysis, in this instance, as noted earlier, is that of multicollinearity or mutual linear dependence among exogenous variables. Multicollinearity, other than where perfect correlation exists is a relative not an absolute problem. The question is when does it become a critical concern for the model. And when so, how should it be treated.

An initial test for the presence of multicollinearity was undertaken using the tolerance of each variable as a measure of potential multicollinearity. The tolerance is that proportion of the variation in a given independent variable not explained by a regression using all other independent variables (i.e., $1 - R^2$ of the regression of X_1 with $X_2 \ldots X_p$). A low tolerance represents high multicollinearity, and vice versa.

Table 13 depicts the tolerance proportions for each independent variable, restated from Table 7. Here, it will be noted that all variables, but car miles and yard hours switching particularly, have a relatively low tolerance. From Table 2 it will be recalled that a simple correlation of .866 existed between CM and YARDHR, thus explaining part of the problem.

A second test for multicollinearity tracks the departure of the correlation matrix of independent variables from orthogonality. If the variables are uncorrelated, the matrix will assume orthogonality. If the variables are highly inter-correlated, a significant departure from orthogonality will occur.

TABLE 13. TOLERANCE VALUES FOR INDEPENDENT VARIABLES.

Variable	Tolerance
Car Miles	.091
Unit Train Miles	.247
Yard Switching Hours	.170

The procedure consists of taking the determinant of the correlation matrix. Since the correlation matrix is a "normalized positive definite matrix", with all elements lying between -1 and +1, the determinant of the matrix itself will assume a value between 0 and one.[7] If the determinant (DET) approaches zero, this connotes

[7] Schilderinck, J.H.F. Regression and Factor Analysis in Econometrics, International Graphics Dordrecht, Leinden, The Netherlands, 1977.

a departure from orthogonality. If the DET approaches one, departure from orthogonality is minimal or nill.[8]

In the instance of the unit train model, the DET of the correlation matrix of independent variables is .0465, connoting a marked departure from orthogonality. This was to be expected and points-out the need for adjusting the cost model.

One obvious solution to multicollinercity is to drop one of the two highly intercorrelated variables. If this does not drastically alter the explanatory benefit of the model, this may, in fact, be the most straightforward and acceptable approach.

Table 14 depcits the results of a regression using such a reformulated model, containing only UTM and CM as exogenous variables. As the Table indicates reformulation of the model reduced somewhat the proportion of the variation in TOTAL explained by the model. The model, however, still explains nearly 98 percent of the variation in total cost. Furthermore, the parameter estimates and the parametric tests for significance were not substantially altered.

TABLE 14. RESULTS OF REFORMULATED REGRESSION MODEL.

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	2	2.49566E+19	1.24783E+19	1233.625	0.0001
ERROR	51	5.15872E+17	1.01151E+16		
C TOTAL	53	2.54724E+19			

ROOT MSE	100574011	R-SQUARE	0.9797
DEP MEAN	601489500	ADJ R-SQ	0.9790
C.V.	16.72083		

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > |T|	TOLERANCE
INTERCEP	1	57748913	18163781	3.179	0.0025	.
CM	1	0.842354	0.028514	29.542	0.0001	0.272707
UTM	1	-45.542450	10.490807	-4.341	0.0001	0.272707

[8] Ibid.

The key question here, though, is whether or not the reformulation mitigated substantially the effects of multicollinearity. Table 14 indicates that while the tolerance measure between the two variables is relatively low, it is a considerable improvement over the previous model. The DET of the correlation matrix is, in this instance, synonymous with the tolerance (.2727) for the 2X2 matrix. This DET may itself be transformed into a test statistic, with a CHI-SQUARE distribution, which can be used to test the significance of the departure from orthogonality.[9] The test statistic consists of the logarithmic transformation of the matrix determinant, as noted below:[10]

(2) $LOG(DET)\ [(T-1)\ \frac{1}{6}\ (2n + 5)]$

where:

T = number of rows in the correlation matrix

n = sample size

This statistic has a CHI-SQUARE distribution with ½ k degrees of freedom. The C-S value can then be used to accept or reject a null hypothesis that there is no signficant departure from orthogonality, or that the DET of the matrix = 1. In this instance, a C-S statistic of 9.886 was calculated. Referring to the CHI-SQUARE table for K-1 or 1 degree of freedom, it was discovered that at the 99th percentile, the table value of 6.635 is less than the C-S value; the null hypothesis is thus rejected.

---

[9] Ibid., Chapter One.

[10] Ibid.

Conclusions on Multicollinearity

The foregoing analysis has established the fact that multicollinearity is present and that the effect is to cause a significant departune from orthogonality. Multicollinearity, however, is not perfect, but is relatively high. This in itself does not bias the regression coefficients exeept by rounding error. Furthermore, all parameter tests are signficant, so the inflation of the SE has not necessarily resulted in the acceptance of a false H_0.

V. CONCLUSIONS

The objective of the study was partially achieved: a model explaining the effects of unit train output on rail costs was developed. The statistical analysis indicates that when controlling for gross output (car miles) the effect of unit train output is to decrease railroad costs by $45 per train mile.

For predictive purposes, the model appears as follows:

Total = 57748913 + .842354 CM - 45.542 UTM + E.

The presence of multicollinearity, however, is a confounding factor which diminishes the statistical viability of the model. The solution would perhaps be to use a technique such as path analysis to point-out the indirect effects of unit train output on costs through other output variables (i.e., yard switching hours, road hours, etc.). rather than include a unit train variable in the equation.