

Estimating the Cost of Construction of Light Water Reactor Plants Using Multiple Regression Model

Rami H. Alamoudi¹

¹(Department of Industrial Engineering, King Abdulaziz University, Jeddah, Saudi Arabia)

Abstract— This paper is an approach to estimate the cost of construction of light water reactor (LWR) plants using information about previous research. A number of variables that are expected to be effective in predicting this cost have been selected. The multiple regression models for the estimation of the cost of construction of light water reactor (LWR) plants using the set of predictors have been developed. The nuclear data frame has 32 rows and 11 columns which mean that we have 10 independent variables and 32 samples. The outcome of the regression analysis shows that the Cost of a future power plant can be estimated using the following independent variables: The date on which the construction permit was issued. The data are measured in years since January 1 1990 to the nearest month. The net capacity of the power plant (MWe). The location of the power plant in the US (If the Plant was constructed in the north-east region of USA). In the end of the paper the equation of estimating the cost of a future power plant has been developed.

Keywords – Light Water Reactor, Matrix Plot, Multiple Regressions, Outlier Test, Stepwise Regression,

I. INTRODUCTION

The search for an alternative source of power other than the fossil based fuels; not only for economical and political reasons but also for environmental reasons; have increased the needs to Construct light water reactor (LWR) plants in the US [1]. The costs related to the construction of light water reactor (LWR) plants started getting a great deal of attention in the late 1960's and early 1970's, due to the future plans of construction and renewal of more (LWR) plants throughout the US. It was important to try to predict the capital cost of construction and renewal of more (LWR) plants, and through this process, some questions started to arise, what are the factors affecting the capital cost of construction?, some important factors that might have an effect on the capital cost of construction as follow:

1. **Date factor :**
 - a. The date on which the construction permit was issued
2. **Time Factors :**
 - a. The time between application for and issue of the construction permit
 - b. The time between issue of operating license and construction permit
3. **Capacity Factor :**
 - a. The net capacity of the power plant (MWe).
4. **Location Factors :**
 - a. The prior existence of a LWR plant at the same site.
 - b. If the Plant was constructed in the north-east region of USA
5. **Technical factors :**
 - a. If a cooling tower is present in the plant
6. **Construction factors :**
 - a. If the nuclear steam supply system was manufactured by Babcock-Wilcox.
 - b. The cumulative number of power plants constructed by each architect-engineer.
 - c. If plants are with partial turnkey guarantees. (some of the power plants had partial turnkey guarantees and it is possible that, for these plants, some manufacturers' subsidies may be hidden in the quoted capital costs)

In this paper, we investigate the effect of each of those factors on the overall capital cost of construction and to use their conclusions for future prediction of the costs related to the construction of light water reactor (LWR) plants in the future.

II. RESEARCH METHODOLOGY

Based on their preliminary researches [2,3] on the factors affecting the costs related to the construction of light water reactor (LWR) plants, we have started collecting data related to the construction of 32 light water reactor (LWR) plants constructed in the US in the late 1960's and early 1970's [1]. The nuclear data frame has 32 rows and 11 columns and contains the following columns:

- cost:** The capital cost of construction in millions of dollars adjusted to 1976 base.
- date:** The date on which the construction permit was issued. The data are measured in years since January 1 1990 to the nearest month.
- t1:** The time between application for and issue of the construction permit.
- t2:** The time between issue of operating license and construction permit.
- cap:** The net capacity of the power plant (MWe).
- pr:** A binary variable where 1 indicates the prior existence of a LWR plant at the same site.
- ne:** A binary variable where 1 indicate that plant was constructed in the north-east region of USA
- ct:** A binary variable where 1 indicates the use of a cooling tower in the plant.
- bw:** A binary variable where 1 indicates that the nuclear steam supply system was manufactured by Babcock-Wilcox.
- cum.n:** The cumulative number of power plants constructed by each architect-engineer.
- pt:** A binary variable where 1 indicates those plants with partial turnkey guarantees.

The first step was to try to evaluate the explanatory power of the variables on the Cost, and also try to detect redundant variables or multi co-linearity among the independent variables. This was to reach the optimum valid model with the best predictors of cost.

A preliminary regression analysis was performed to assess all that, through a matrix plot (a plot showing predictors vs predicted and predictors vs one another) and Residual vs fitted values plot [4]. We can notice two problems, problem of non constant variance in the residual plots, so transformation of the predicted Variable (cost) was needed to overcome this problem. And a non linear relationship between Cost and the cumulative number of power plants constructed by an engineer (Cum.n), So we need to standardized the data and try for a transformation of cum.n too and we tried several transforms (monitoring the significance of the partial T-test). The results were b that the best transform to eliminate the non-constant variance was $1/\ln(\text{cost})$ and to include cum.n, cum.n^2 , Cum.n^3 in the model was the only way to get a significant linear relationship for the variable cum.n.

The next Phase was getting rid of redundant variables (like t2; having no relationship with $1/\ln(\text{cost})$) and multicollinearity among variables (like the case between Date and t1) and reducing the variables only the most significant and with the highest explanatory power using stepwise technique [5].

Through regressions and evaluation of the partial t-tests, residual plots and sequential sum of squares and other methods like best subset and stepwise regressions, we ended up with three most significant variables [6,7]. They showed random residual plots, partial t-tests with p-values of zero, reasonable ssq and a variance inflation factor of one. The only problem was two outliers in the residual vs fitted values plot that didn't show in the residual plot of the variables, which suggested the presence of influential outliers, so we performed some outliers tests like cooks distance, hat matrix and DFITS they all showed the same two outliers (they were the same outliers on the residual vs fitted value graph) , so going back to the data sources , we were able to eliminate those two points and we ended up with a reasonable model for predicting the Cost.

III. RESEARCH ANALYSIS

Looking for directly correlated variables, we decided to perform a multiple linear regression run, and check the Matrix plot (Fig. 1), and since the matrix plot doesn't give enough information about binary variables, we only plotted the quantitative variables.

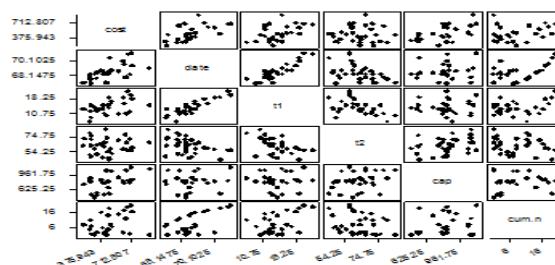


Fig. 1 The matrix plot for the preliminary regression model

The Preliminary Regression Observations will be as follow: The matrix plot and normal probability and residual plots will be discussed as follow:

1- Matrix plot :

a. Predicted vs Predictors :

- i. Cost vs Date : Looks like a Linear relation ..fine
- ii. Cost vs t1 : Looks like a Linear relation ..fine
- iii. Cost vs t2 : Looks completely random ..suggests no relationship
- iv. Cost vs Cap : Looks like a Linear relation ..fine
- v. Cost vs Cum.n : Looks like a non-Linear relation .. (U shaped), might need transformation.

b. Predictors vs Predictors :

- i. Date vs t1 : Looks like a Linear relation...eliminate one of them (later we will see that we chose t1)
- ii. Date vs t2 : Looks like an inverse Linear relation...might eliminate one of them (later we will see that we chose t2)
- iii. Date vs Cap : looks random
- iv. Date vs cum.n : Looks like a Linear relation .

2- Normal probability and residual plots :

a. The normal probability plot looks very good

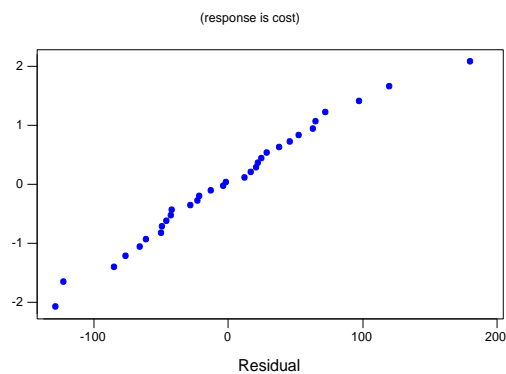


Fig. 2 Normal probability plot of the residual for the preliminary regression model

b. Residuals vs fitted values ...looks like there is some funneling there ...with may be one outlier. So a transformation might be recommended to avoid the non-constant variance (that shows in the funneling).

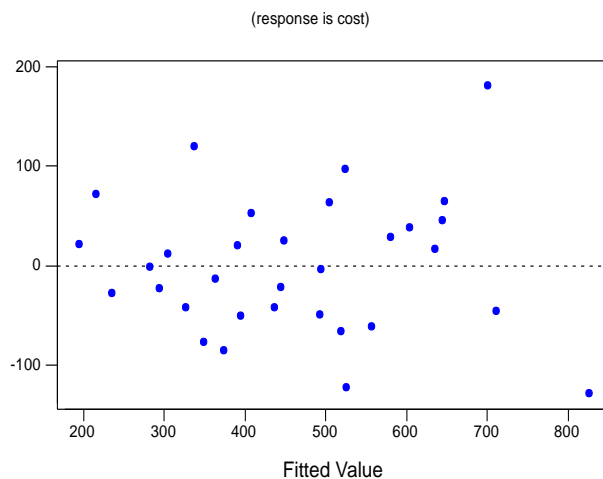


Fig. 3 Residual versus the fitted values for the preliminary regression model

c. Residuals vs Date random with a couple of outliers maybe.

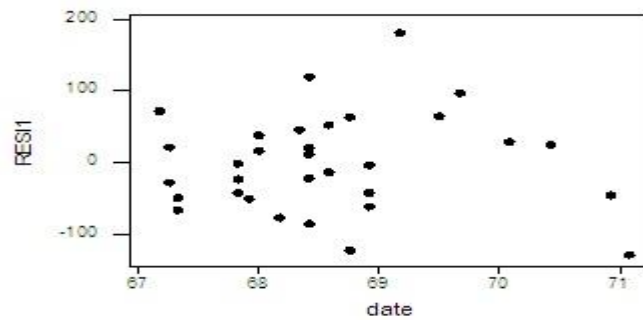


Fig. 4 Residuals vs Date random for the preliminary regression model

d. Residuals vs t1 random

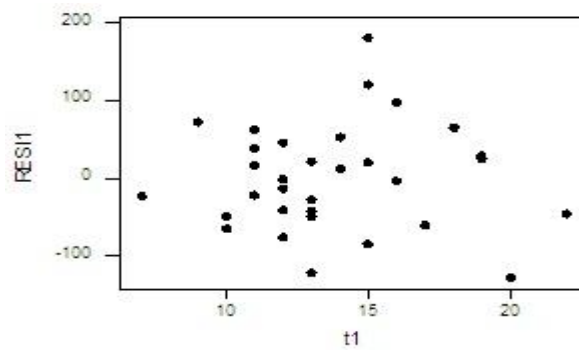


Fig. 5 Residuals vs t1 random for the preliminary regression model

e. Residuals vs t2 random

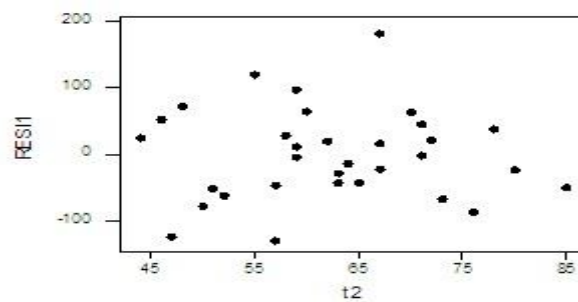


Fig. 6 Residuals vs t2 random for the preliminary regression model

f. Residuals vs Cum.n , “U “ shaped.. support the transformation decision

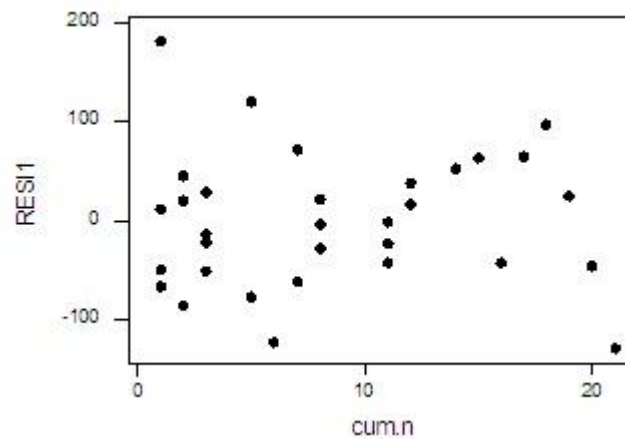


Fig. 7 Residuals vs Cum.n random for the preliminary regression model

From the above analysis, we have non constant variance clear in the funneling of the residual verses fit plot, thus we need to search for a proper transform.

After a number of trials to transform Cost and checking for the randomness of the residual versus fitted value plot, and also looking at the matrix plot, we ended up choosing the $1 / \ln(\text{cost})$, as the best possible transform. [8].

It was noted earlier that the variable cum.n appeared to have a non linear relationship with the cost, this non-linear relationship stayed after transformation but the variable stayed statistically insignificant, so I decided to include the variable cum.n^2 and cum.n^3 and standardize but still. So we included also cum.n^2 and cum.n^3 and standardized and now we got them significant [9]

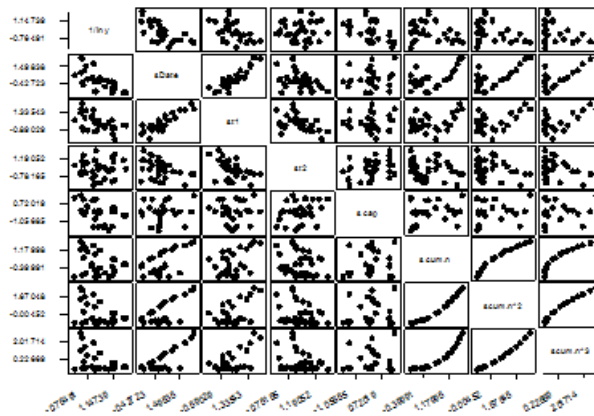


Fig. 8 The matrix plot after transformation

Thus, reapplying the matrix plot and Normal probability and residual plots as follow:

1. Matrix plot :

Predicted vs Predictors :

- I. $s(1/\ln\text{cost})$ vs sDate : Looks like a –ve Linear relation ..fine
- II. $s(1/\ln\text{cost})$ vs st1 : Looks like a –ve Linear relation ..fine
- III. $s(1/\ln\text{cost})$ vs st2 : Looks completely random ..suggests no relationship
- IV. $s(1/\ln\text{cost})$ vs sCap : Looks like a –ve Linear relation ..fine
- V. $s(1/\ln\text{cost})$ vs Cum.n cum.n^2 and cum.n^3 : Looks like a –ve Linear relation ..fine

Predictors vs Predictors :

- VI. sDate vs st1 : Looks like a Linear relation...eliminate one of them (I Choose st1)
- VII. sDate vs st2 : Looks like an inverse Linear relation...might eliminate one of them
- VIII. sDate vs sCap : looks random
- IX. sDate vs sCum.n scum.n^2 and scum.n^3 : Looks like a Linear relations there .
- X. st1 vs sCum.n scum.n^2 and scum.n^3 : Looks like a Linear relations there

The Normal probability and residual plots (After Transformation) will be as shown in Fig. 9.

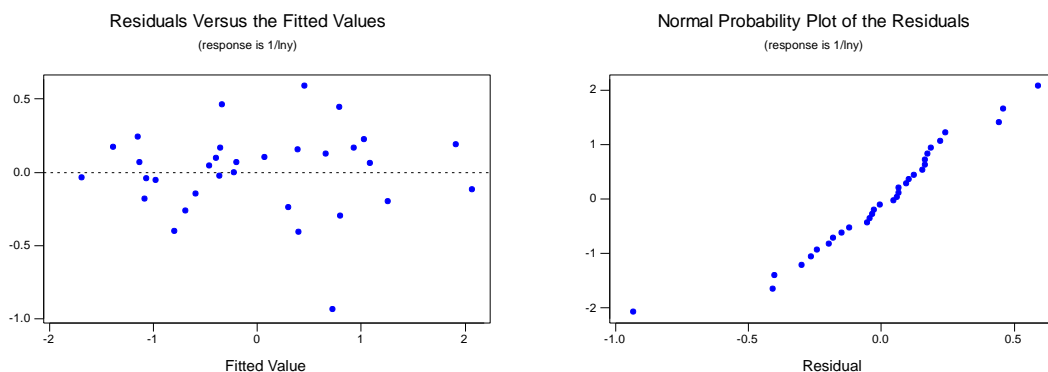


Fig. 9 Normal probability and residual vs fitted values plots after transformation

The normal probability plot has an obvious outlier there. Residuals vs fitted values ...looks like there is no funneling now...but there might be one outlier. The Regression Analysis after Transformation will be as follow:

The regression equation is:

$$1/\ln y = -0.0000 - 1.02 \text{ sDate} + 0.042 \text{ st1} + 0.009 \text{ st2} - 0.481 \text{ scap} + 0.0687 \text{ spr} \\ - 0.295 \text{ sne} - 0.263 \text{ sct} + 0.0716 \text{ sbw} + 3.18 \text{ scum.n} - 7.80 \text{ scum.n}^2 \\ + 5.27 \text{ scum.n}^3$$

Predictor	Coef	StDev	T	P
Constant	-0.00000	0.06359	-0.00	1.000
sDate	-1.0230	0.1761	-5.81	0.000
st1	0.0421	0.1573	0.27	0.792
st2	0.0087	0.1276	0.07	0.946
scap	-0.48129	0.07592	-6.34	0.000
spr	0.06866	0.09324	0.74	0.470
sne	-0.29468	0.06942	-4.24	0.000
sct	-0.26255	0.07984	-3.29	0.004
sbw	0.07159	0.08800	0.81	0.426
scum.n	3.1803	0.8867	3.59	0.002
scum.n ²	-7.795	2.021	-3.86	0.001
scum.n ³	5.267	1.246	4.23	0.000

S = 0.3597 R-Sq = 91.7% R-Sq(adj) = 87.1%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	11	28.4120	2.5829	19.96	0.000
Residual Error	20	2.5880	0.1294		
Total	31	31.0000			

Source	DF	Seq SS
sDate	1	12.2940
st1	1	1.0097
st2	1	1.0150
scap	1	3.7406
spr	1	1.0806
sne	1	3.5404
sct	1	0.9954
sbw	1	0.3975
scum.n	1	1.8744
scum.n ²	1	0.1515
scum.n ³	1	2.3129

Unusual Observations

Obs	sDate	1/lny	Fit	StDev Fit	Residual	St Resid
9	-0.16	-0.2090	0.7247	0.1729	-0.9337	-2.96R
18	-0.16	1.0483	0.4590	0.2235	0.5893	2.09R

We can observe the following:

- Over All F-test :
 H₀ : all B_i = zero
 H_a : at least one B_i not = zero
 p-value = 0, Reject the null hypothesis
 Therefore at least one B_i not equal to zero. The model stands.
- Predictors :

Best Subsets Regression: 1/lny versus sdate, st1, ...

Response is 1/lny

Vars	R-Sq	R-Sq(adj)	C-p	S	e	1	2	p	r	e	t	w	n	2	3
1	39.7	37.6	116.6	0.0082526	X										
1	20.0	17.4	163.6	0.0095008	X										
2	57.0	54.0	77.1	0.0070874	X	X									
2	50.1	46.7	93.5	0.0076304	X		X								
3	67.7	64.2	53.4	0.0062511	X	X	X								
3	65.0	61.2	59.9	0.0065094	X	X	X								
4	74.5	70.7	39.1	0.0056541	X	X	X	X							
4	73.8	69.9	40.9	0.0057370	X	X	X	X							
5	81.8	78.4	23.5	0.0048621	X	X	X	X	X						
5	81.3	77.7	24.7	0.0049316	X	X	X	X	X						
6	85.6	82.1	16.6	0.0044223	X	X	X	X	X	X					
6	83.0	78.9	22.8	0.0048008	X	X	X	X	X	X					
7	90.8	88.1	6.1	0.0036098	X	X	X	X	X	X					
7	86.4	82.4	16.7	0.0043869	X	X	X	X	X	X					
8	91.3	88.3	6.8	0.0035748	X	X	X	X	X	X	X				
8	91.2	88.1	7.1	0.0036023	X	X	X	X	X	X	X				
9	91.6	88.2	8.1	0.0035909	X	X	X	X	X	X	X	X			
9	91.4	87.8	8.7	0.0036433	X	X	X	X	X	X	X	X			
10	91.6	87.7	10.0	0.0036693	X	X	X	X	X	X	X	X	X		
10	91.6	87.6	10.1	0.0036754	X	X	X	X	X	X	X	X	X		
11	91.7	87.1	12.0	0.0037595	X	X	X	X	X	X	X	X	X	X	

Stepwise Regression: 1/lny versus sdate, st1, ...

Alpha-to-Enter: 0.15 Alpha-to-Remove: 0.15

Response is 1/lny on 11 predictors, with N = 32

Step	1	2	3	4	5
Constant	0.6100	0.6232	0.6026	0.7233	0.7423
sdate	-0.0065	-0.0064	-0.0061	-0.0078	-0.0081
T-Value	-4.44	-5.10	-5.46	-6.52	-7.69
P-Value	0.000	0.000	0.000	0.000	0.000
scap		-0.00002	-0.00002	-0.00003	-0.00003
T-Value		-3.42	-3.90	-4.80	-5.45
P-Value		0.002	0.001	0.000	0.000
sne			-0.0078	-0.0090	-0.0084
T-Value			-3.05	-3.82	-4.03
P-Value			0.005	0.001	0.000
scum.n				0.00054	0.00057
T-Value				2.69	3.25
P-Value				0.012	0.003
sct					-0.0055
T-Value					-3.08

P-Value	0.005				
S	0.00825	0.00709	0.00625	0.00565	0.00493
R-Sq	39.66	56.98	67.69	74.51	81.32
R-Sq(adj)	37.65	54.01	64.22	70.73	77.73
Mallows C-p	116.6	77.1	53.4	39.1	24.7

Looking at the Stepwise Regression, Best Subsets Regression and the partial T-test P-value of each predictor, and its corresponding seq ss [9]. We can say that:

- sdate, scap, sne, sCum.n scumn^2 and scumn^3 are all significant with p-value < 0.05 But
- since the matrix plot showed linear relation between Sdate and sCum.n scumn^2 and scumn^3, And also linear relation between Sdate and st1, we will try eliminating st1, sCum.n, scumn^2 and scumn^3 from my model.
- st2 with p-value > 0.05 and no relationship with s (1/ln(cost)), so we decided to eliminate it from the model.

The run after First Reduction, after eliminating st1, scum.n, scumn^2 and scumn^3 and st2 we get the following:

The regression equation is

$$1/\ln y = -0.0000 - 0.590 \text{ sDate} - 0.432 \text{ scap} + 0.090 \text{ spr} - 0.292 \text{ sne} - 0.231 \text{ sct} + 0.050 \text{ sbw}$$

Predictor	Coef	StDev	T	P
Constant	-0.00000	0.09893	-0.00	1.000
sDate	-0.5905	0.1027	-5.75	0.000
scap	-0.4322	0.1028	-4.20	0.000
spr	0.0903	0.1035	0.87	0.391
sne	-0.2917	0.1021	-2.86	0.008
sct	-0.2314	0.1028	-2.25	0.033
sbw	0.0498	0.1032	0.48	0.634

S = 0.5596 R-Sq = 74.7% R-Sq(adj) = 68.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	6	23.1698	3.8616	12.33	0.000
Residual Error	25	7.8302	0.3132		
Total	31	31.0000			

Source	DF	Seq SS
sDate	1	12.2940
scap	1	5.3694
spr	1	0.6583
sne	1	3.1249
sct	1	1.6502
sbw	1	0.0729

Unusual Observations

Obs	sDate	1/lny	Fit	StDev Fit	Residual	St Resid
19	0.58	-1.7211	-0.6758	0.2409	-1.0452	-2.07R
26	2.46	-1.2160	-2.3416	0.3545	1.1256	2.60R

Best Subsets Regression: 1/lny versus sdate, scap, spr, sne, sct, sbw
Response is 1/lny

s
d s
a c s s s s

Vars	R-Sq	Mallows R-Sq(adj)	C-p	t a p n c b	S e p r e t w
1	39.7	37.6	31.7	0.0082526	X
1	18.4	15.6	52.8	0.0095993	X
2	57.0	54.0	16.6	0.0070874	X X
2	50.1	46.7	23.4	0.0076304	X X
3	67.7	64.2	8.0	0.0062511	X X X
3	65.0	61.2	10.7	0.0065094	X X X
4	73.8	69.9	4.0	0.0057370	X X X X
4	69.2	64.6	8.5	0.0062167	X X X X
5	74.5	69.6	5.2	0.0057620	X X X X X
5	74.0	69.0	5.8	0.0058222	X X X X X
6	74.7	68.7	7.0	0.0058490	X X X X X X

Stepwise Regression: 1/lny versus sdate, scap, spr, sne, sct, sbw

Alpha-to-Enter: 0.15 Alpha-to-Remove: 0.15
Response is 1/lny on 6 predictors, with N = 32

Step	1	2	3	4
Constant	0.6100	0.6232	0.6026	0.6143
sdate	-0.0065	-0.0064	-0.0061	-0.0062
T-Value	-4.44	-5.10	-5.46	-6.08
P-Value	0.000	0.000	0.000	0.000
scap		-0.00002	-0.00002	-0.00002
T-Value		-3.42	-3.90	-4.17
P-Value		0.002	0.001	0.000
sne			-0.0078	-0.0071
T-Value			-3.05	-3.01
P-Value			0.005	0.006
sct				-0.0052
T-Value				-2.50
P-Value				0.019
S	0.00825	0.00709	0.00625	0.00574
R-Sq	39.66	56.98	67.69	73.76
R-Sq(adj)	37.65	54.01	64.22	69.87
Mallows C-p	31.7	16.6	8.0	4.0

Looking at the Stepwise Regression, Best Subsets Regression and the partial T-test P-value of each predictor, and its corresponding seq ss. We can see that spr, sct, sbw, and spt are all with p-value > 0.05 and with small contribution in the seq ss. So we decide to REMOVE ct, bw, pr and pt, from the model, and run it again. The run after second Reduction of variables [10], now we ran only the variables Date, Cap and ne against the 1/ln(cost)

The regression equation is
1/ln y = 0.603 - 0.00607 date - 0.000023 cap - 0.00781 ne

Predictor	Coef	StDev	T	P	VIF
Constant	0.60264	0.07622	7.91	0.000	
date	-0.006068	0.001111	-5.46	0.000	1.0
cap	-0.00002313	0.00000593	-3.90	0.001	1.0
ne	-0.007811	0.002564	-3.05	0.005	1.0

S = 0.006251 R-Sq = 67.7% R-Sq(adj) = 64.2%

Analysis of Variance					
Source	DF	SS	MS	F	P
Regression	3	0.00229188	0.00076396	19.55	0.000
Residual Error	28	0.00109413	0.00003908		
Total	31	0.00338600			

Best Subsets Regression

Response is 1/ln y

Vars	R-Sq	Adj. R-Sq	C-p	Max	Delta
1	39.7	37.6	24.3	0.0082526	X
1	18.4	15.6	42.7	0.0095993	X
2	57.0	54.0	11.3	0.0070874	X X
2	50.1	46.7	17.2	0.0076304	X X
3	67.7	64.2	4.0	0.0062511	X X X

Stepwise Regression: 1/lny versus sdate, scap, sne

Alpha-to-Enter: 0.15 Alpha-to-Remove: 0.15

Response is 1/lny on 3 predictors, with N = 32

Step	1	2	3
Constant	0.6100	0.6232	0.6026
sdate	-0.0065	-0.0064	-0.0061
T-Value	-4.44	-5.10	-5.46
P-Value	0.000	0.000	0.000
scap		-0.00002	-0.00002
T-Value		-3.42	-3.90
P-Value		0.002	0.001
sne			-0.0078
T-Value			-3.05
P-Value			0.005
S	0.00825	0.00709	0.00625
R-Sq	39.66	56.98	67.69
R-Sq(adj)	37.65	54.01	64.22
Mallows C-p	24.3	11.3	4.0

IV. DISCUSSION

After conducting the above regression methods and the stepwise regressions we can conclude the following results:

A. Over All F-test :

H0 : all Bi = zero

Ha : at least one Bi not = zero

p-value = 0, Reject the null hypothesis

therefore at least one Bi not equal to zero . the model stands.

B. Predictors :

Looking at Stepwise Regression, Best Subsets Regression and the P-value of each predictor, and its corresponding seq ss.

We can say that ;

All p-values are ≈ zero....so we are good.

C. Best Subsets :

Looks like we won't be able to get rid of any of them , or else bias problems will occur. BUT there are 2 obvious outliers in the residual plot... so we need to run the test for outlier.

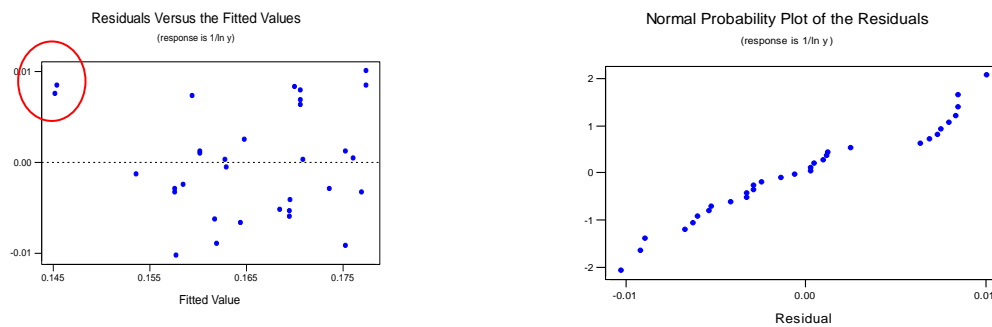


Fig. 10 Normal probability and residual vs fitted values plots after factors elimination

Therefore, from Fig 10 we have to conduct the outlier tests to see whether these points should be eliminated or not. The outlier test will be as follow:

1. **Hat matrix indicator**

Index = $2p/n = 0.1875$

Therefore this subjects that cases 22 and 26 are highly influential

2. **Cook's Distance :**

The cases 22 and 26 have high cook's distance, this supports that they are highly influential

3. **DFFITs :**

Index = $2 \sqrt{p/n} = 0.61$

The cases 22 and 26 also have high DFFITS , this supports that they are highly influential

4. **DFBitas Appendix (C):**

The cases 22 and 26 also have high Effect on the parameters b_0 , b_1 , and b_3 .

When we went back to the data collected it showed that the two nuclear power stations had different characteristics than the rest and that it would be wise to remove them from the data set.

The Model after Removing case 26:

Table 1: Outlier test

	Hat Mat	COOKS	DFITS
1	0.142463	0.000109	0.02048
2	0.140142	0.042856	-0.41443
3	0.140142	0.034836	-0.37229
4	0.196627	0.020828	-0.28517
5	0.196627	0.016356	-0.25238
6	0.231088	0.00021	0.02843
7	0.045655	0.022448	0.30464
8	0.164355	0.016177	-0.25128
9	0.042027	0.007902	-0.17687
10	0.131365	0.088482	-0.61026
11	0.105586	0.007065	-0.16579
12	0.125995	0.05682	0.48191
13	0.120704	0.083456	-0.59373
14	0.091118	0.00069	0.05161
15	0.047048	0.002109	0.09046
16	0.044061	0.00528	-0.14389
17	0.143094	0.002175	-0.09168
18	0.120704	0.001423	0.07413
19	0.116337	0.100083	-0.65808
20	0.091118	0.001079	0.06456
21	0.049932	0.015735	-0.2518
22	0.273792	0.238601	1.00588
23	0.084868	0.025216	-0.31811
24	0.118175	0.005791	-0.14991
25	0.233622	0.000818	-0.05617
26	0.325807	0.261149	1.04478
27	0.098258	0.078782	0.58213
28	0.099677	0.000171	0.02568
29	0.060452	0.017759	0.26704
30	0.060452	0.02067	0.28907
31	0.098258	0.055546	0.4807
32	0.060452	0.027888	0.33862

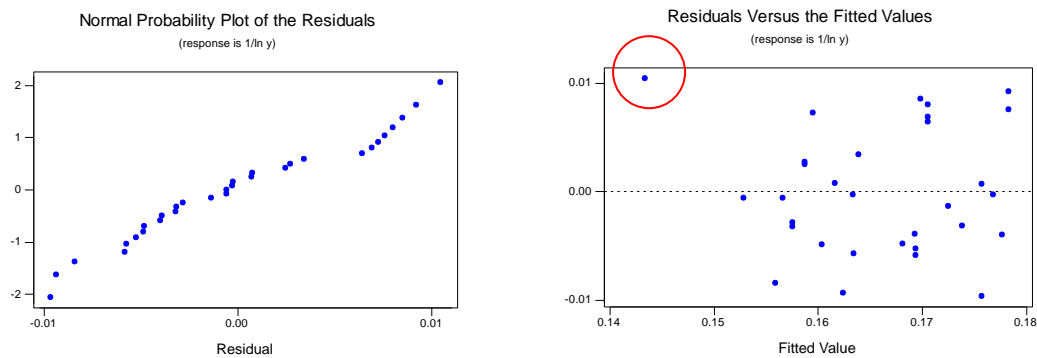


Fig. 11 Normal probability and residual vs fitted values plots after eliminating case 22

We can see that we actually have a better normal probability plot, and we only have one outlier now in the residual plot. Note that the model is still significant. The Model after Removing case 22 will be as shown in Fig. 11 We can see that we actually have a better normal probability plot, and we only have no outliers in the residual plot. Note that the model is still significant, and with even more overall F-statistic value and more R-sq (adj)

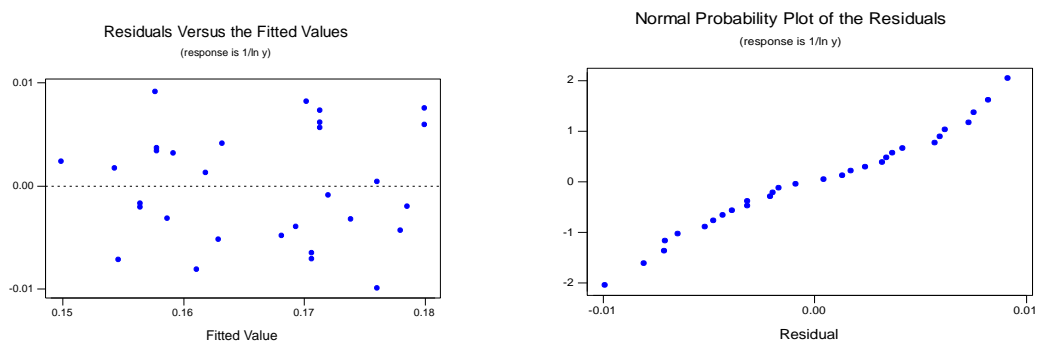


Fig. 12 Normal probability and residual vs fitted values plots after eliminating case 22 and 26

And therefore, after removing cases 22 and 26 the regression equation will be as follow:

$$1/\ln y = 0.760 - 0.00833 \text{ date} - 0.000027 \text{ cap} - 0.00860 \text{ ne}$$

Predictor	Coef	StDev	T	P	VIF
Constant	0.75995	0.09067	8.38	0.000	
date	-0.008325	0.001313	-6.34	0.000	1.0
cap	-0.00002709	0.00000575	-4.71	0.000	1.0
ne	-0.008596	0.002469	-3.48	0.002	1.0

$$S = 0.005718 \quad R\text{-Sq} = 72.3\% \quad R\text{-Sq}(\text{adj}) = 69.1\%$$

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	0.00221929	0.00073976	22.63	0.000
Residual Error	26	0.00085008	0.00003270		
Total	29	0.00306937			

Source	DF	Seq SS
date	1	0.00106928
cap	1	0.00075374
ne	1	0.00039628

V. SUMMARY AND RESULTS

From the above we can conclude that the cost of a future power plant can be estimated using,

- **date:** The date on which the construction permit was issued. The data are measured in years since January 1 1990 to the nearest month.
- **cap:** The net capacity of the power plant (MWe).
- **ne:** A binary variable where 1 indicate that plant was constructed in the north-east region of US.

And this can be done using the regression equation

$$1/\ln \text{ Cost} = 0.760 - 0.00833 \text{ date} - 0.000027 \text{ cap} - 0.00860 \text{ ne}$$

Or in other words, the Cost of a future power plant can be calculated from the following relationship:

$$\text{Cost} = \exp (1 / (0.760 - 0.00833 \text{ date} - 0.000027 \text{ cap} - 0.00860 \text{ ne}))$$

VI. CONCLUSIONS

In this paper the multiple regression models to estimate the cost of constructing Light Water Reactor (LWR) plants has been developed using previous data. Since the goal would be minimizing the cost of construction, and through the model we see that the cost would increase with the increase of Date, Capacity and the location (in the North-East region of the US). So based on our model we can recommend that the construction permit should be issued as soon as possible, the net capacity of the power plant should be kept with the expected loads and not to construct in the North-East region of the US. In the future, the proposed regression model can be enhanced by considering other factors that might affect the cost of construction. Another promising avenue for future research is to apply the proposed regression model to similar application.

VII. REFERENCES

- [1] Cox, D.R and Snell, E.J. (1981) *Applied Statistics: Principles and Examples*. Chapman and Hall (Source of the data)
- [2] B.R. Sehgal, "Stabilization and termination of severe accidents in LWRs", *Nucl. Eng. Design* 236, 1941 (2006).
- [3] B R Sehgal, Light Water Rector (LWR) Safety, Nuclear Engineering and Technology, VOL.38 NO.8, 2006
- [4] Babyak MA. What you See May not be What you Get: A Brief, Nontechnical Introduction to Over Fitting in Regression-Type Models. *Psychosom Med* 2004;66:411–21.
- [5] Derksen S, Keselman H. Backward, Forward and Stepwise Automated Subset Selection Algorithms: Frequency of Obtaining Authentic and Noise Variables, *Br J Math Stat Psychol* 1992;45:265–82.
- [6] Steyerberg EW, Eijkemans MJ, Harrell FE Jr., et al. Prognostic Modeling with Logistic Regression Analysis: in Search of a Sensible Strategy in Small Data Sets. *Med Decis Making* 2001;21:45–56.
- [7] Kleinbaum D, Klein M. *Logistic regression*. 2nd edn. New York: Springer-Verlag, 2002.
- [8] S Abdul-Wahab, Charles S. Bakheit, Saleh M. Al-Alawi , Principal Component and Multiple Regression Analysis in Modeling of Ground-Level Ozone and Factors Affecting its Concentrations, *Environmental Modeling & Software*, Volume 20, Issue 10, October 2005, Pages 1263-1271
- [9] H Cordell and D Clayton, A Unified Stepwise Regression Procedure for Evaluating the Relative Effects of Polymorphisms within a Gene Using Case/Control or Family Data: Application to HLA in Type 1 Diabetes, *AJHG*, V 70, Issue 1 2002, 124-141
- [10] Neter, Kutner, Nachtsheim, Wasserman *Applied Linear Regression Models*, 3rd Edition. 1996